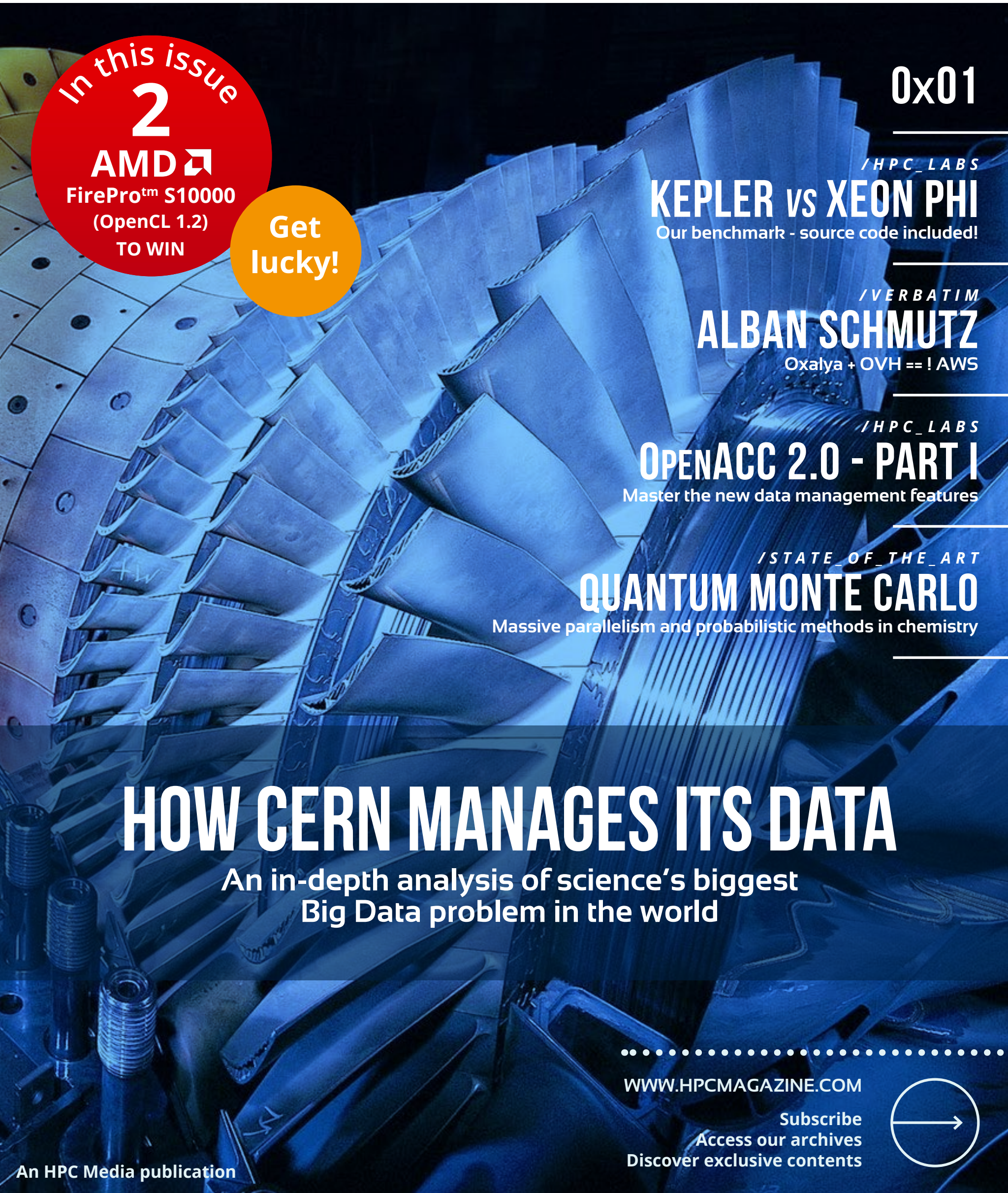
JANUARY 2014

HighPerformanceComputing

The reference media in HPC and Big Data

Volume I | Number 1

Americas



In this issue

2

AMD FirePro™ S10000 (OpenCL 1.2) TO WIN

Get lucky!

0x01

/HPC_LABS

KEPLER vs XEON PHI

Our benchmark - source code included!

/VERBATIM

ALBAN SCHMUTZ

Oxalya + OVH == ! AWS

/HPC_LABS

OPENACC 2.0 - PART I

Master the new data management features

/STATE_OF_THE_ART

QUANTUM MONTE CARLO

Massive parallelism and probabilistic methods in chemistry

HOW CERN MANAGES ITS DATA

An in-depth analysis of science's biggest Big Data problem in the world

WWW.HPCMAGAZINE.COM

Subscribe
Access our archives
Discover exclusive contents

An HPC Media publication

Volume I, number 1

Publisher
Frédéric Milliot

Executive Editors
Stéphane Bihan
Eric Tenin

Contributing Editors
Nick Anderson
Michel Caffarel
Stéphane Chauveau
Amaury de Cizancourt
Steve Conway
Florent Duguet
Wolfgang Gentzsch
Alex Roussel
Anthony Scemama

Advisory Board
(to be announced)

Communication
Laetitia Paris

Submissions
We welcome submissions. Articles must be original and are subject to editing for style and clarity.

Contacts
editorial@hpcmagazine.com
subscriptions@hpcmagazine.com
advertising@hpcmagazine.com

HighPerformanceComputing
HPC MEDIA
11, rue du Mont Valérien
92210 St-Cloud - France

From the Publisher

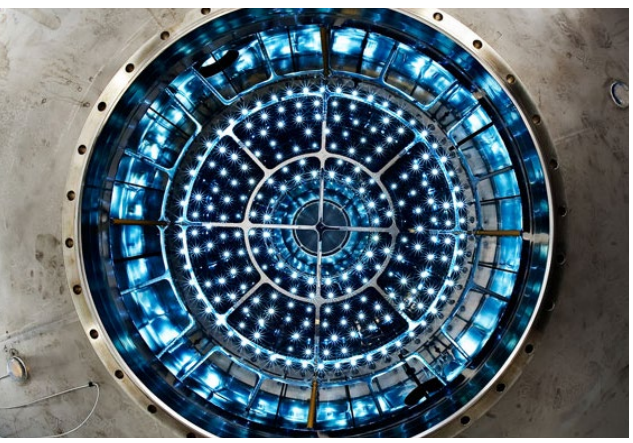
Welcome...

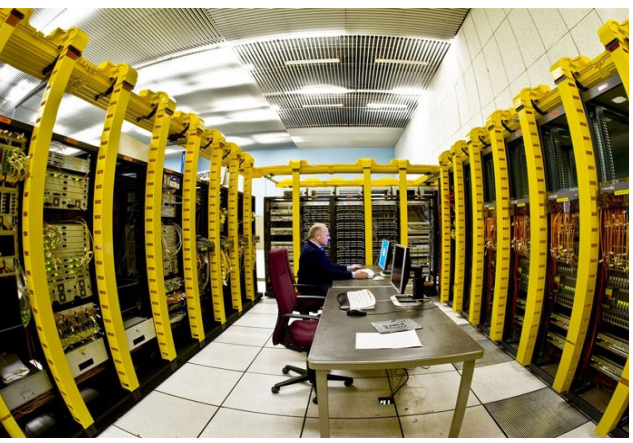
Welcome to the first issue of **HighPerformanceComputing**, the new digital media fully dedicated to HPC and Big Data news, technologies, uses and research. "HPC Mag" (as we call it internally) has been designed with you in mind. We wanted to create a balanced media that appeals to practicing users, developers and researchers, both in the academic and private sectors, but also to new entrants in the field of extreme and technical computing. That is why, along with in-depth, mostly peer-reviewed articles coming from our network of international contributors, we'll have at least one evangelizing "success story" every month. About contributions, you should consider becoming an author. If you feel like sharing your thoughts and discoveries with the community, don't hesitate to submit a paper, a report, a column. We'll work together with you to give it the audience it deserves.

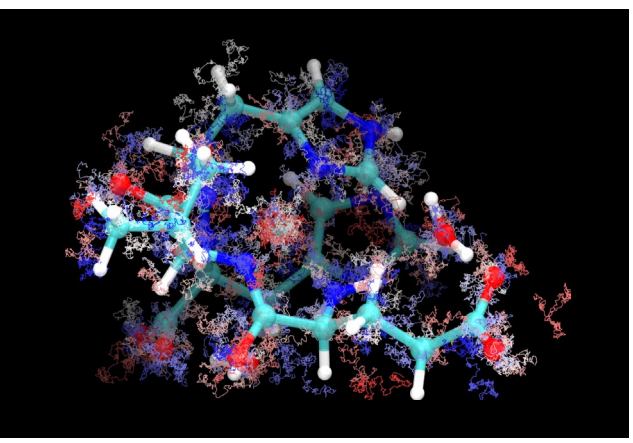
Although this first issue is far from being perfect, we hope **HighPerformanceComputing**, along with its website, TV channel and the forthcoming HPC Labs, will soon become an essential part of your professional life, a resource that you depend on to keep up with our constantly evolving science and practice. Needless to say, we welcome all comments and suggestions. To make a long story short, we look forward to growing with you.

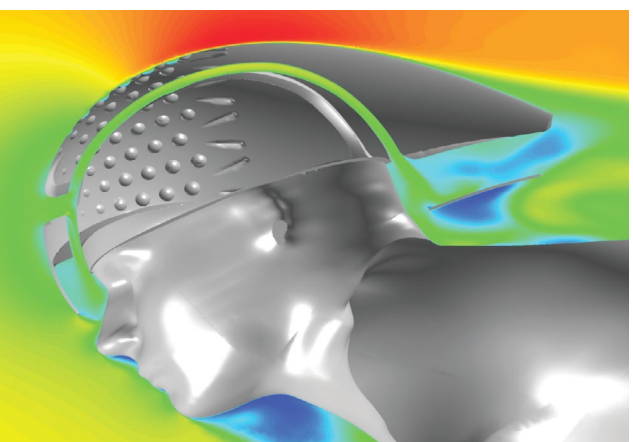
Happy reading!

frederic@hpcmagazine.com









CONTENTS

JANUARY 2014

Visit www.hpcmagazine.com for multimedia content related to the features in this issue.

CRAY®

Cray® CS300™ Cluster Supercomputers for Big Computing and Big Data Challenges

/NEWS

Argonne to *really* standardize MPI

It may sound surprising considering that MPI is a *de facto* standard but MPI implementations are actually far from standardized among vendors. Therefore, with every new release, developers have to carefully test and fine-tune their codes using the latest libraries. A definitely time-consuming process that may soon belong to the past.

Researchers at Argonne Laboratory have just established an initiative that promises to maintain runtime compatibility between implementations regardless - so they say - of the underlying hardware architectures. The idea is to develop an Application Binary Interface that would standardize such details as how functions are called or the size, layout and alignment of datatypes, resulting in a set of runtime conventions. The good news is that the project, called MPICH ABI Compatibility Initiative, involves several prominent producers, including Cray, IBM and Intel.

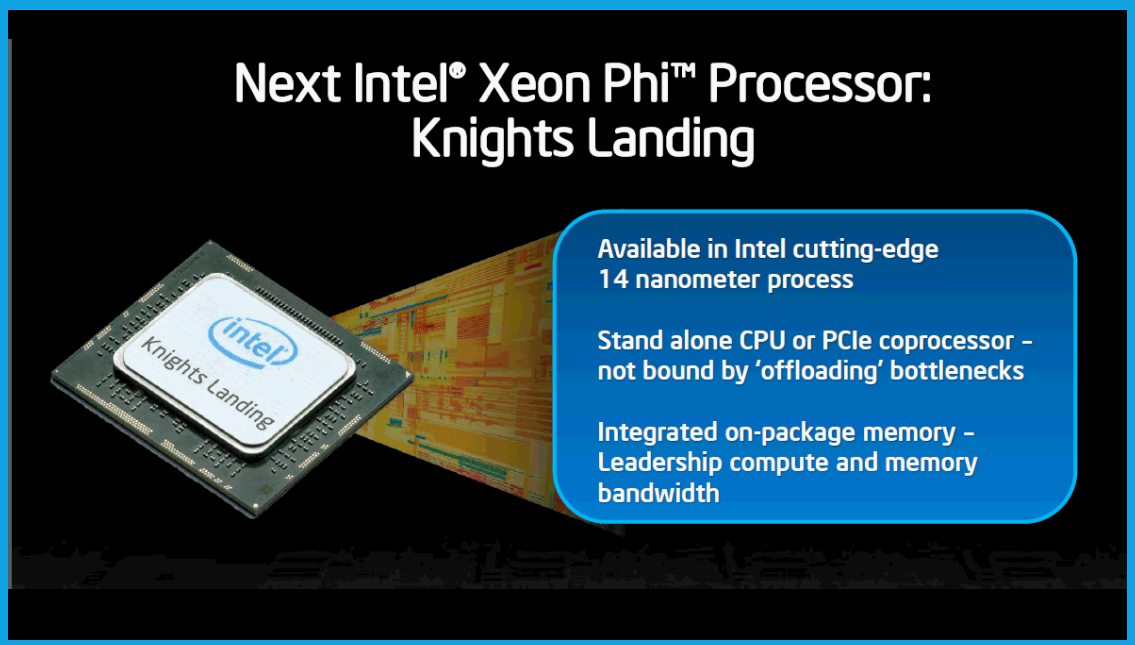
"*One of the primary aims of this initiative is for all parties to agree on a schedule for necessary ABI changes*" said Michael Blocksome, who oversees the development of the MPI software implementation at IBM. "*This initiative will provide a binary-compatible upgrade path between the MPI-2 and MPI-3 standards, which will make customers' MPI-3 transitions easier and more successful. It will also allow our binary compatibility to extend across our collaborators' MPI implementations, which is important both to our customers and to our vendor partners*" said Dave Poulsen, who leads Intel's MPI and Parallel Runtimes team. This project is not limited to just MPICH-derived implementations. Other collaborators are officially welcome to join and participate in the initiative for the benefit of users, application developers and system administrators alike. In the meantime, the first packages are announced for release before this Summer.

Seagate Technology to Acquire Xyratex

Seagate Technology and Xyratex announced that they have entered into a definitive agreement under which Seagate will acquire all outstanding shares of Xyratex in a transaction valued at approximately $374 million. Although not a game-changer, this acquisition will bring Western Digital's arch-rival some vertical integration in manufacturing processes (Xyratex's proprietary test equipments are used for testing hard drives before they are shipped to customers) and expand its HPC storage solutions portfolio. No wonder, then, that Seagate intends to maintain the ClusterStor business unit as a separate entity and continue to support Xyratex's numerous OEM customers. The latest addition to the line, ClusterStor 9000, was officially announced a few days before SC'13, featuring a 60 GB/s throughput per rack.

XEON PHI 2.0: INTEL SAYS MORE

One year after the official announcement of Xeon Phi 1.0 ("Knight's Corner"), Intel took advantage of the SuperComputing Conference to organize a round-table discussion about the future of its x86 accelerator. This was the occasion for the company to publicly state its long–term intentions and to remind the audience that the objective of Xeon Phi is to provide a non-disruptive transition path toward multicore and manycore architectures, i.e. to allow applications to be executed on an ever larger number of processors within the context of x86 conventional programming.



As developers know too well, the main problem of heterogeneous computing today is data loading onto the accelerators. Intel intends to provide a lasting solution with the forthcoming "Knight's Landing" memory architecture: a cache mechanism located within the core itself, and in-package high bandwidth memory to be connected to large quantities of DDR4. The use of this cache to control memory placement and exploit the extra cores should allow Knight's Landing - according to Intel engineers - to run applications natively rather than in accelerated mode, thus avoiding data transfers.

As a bonus, we now know more about the main specifications of Phi 2.0: manufactured in 14 nm technology and available in socket format, it will carry 72 Airmont (Intel Atom) cores with out-of-order 512-bit vector instructions (AVX-512), 8 to 16 GB of memory within the accelerator, 6 access channels to the future DDR4 memory (up to 384 GB @ 115 GB/s!) and 36 PCIe Gen3 lines. On the performance front, Intel announces 3 Tflops double precision, or 3X Knight's Corner performance for a 160 to 200 W TDP. The card version will be limited to two channels of DDR4 for a maximum of 64 GB of RAM. An "F" version should also be available, with optical interconnects (the "F" is for Fabric) and a 15 W higher TDP. Regarding the release date, nothing has changed: depending on the version, you will need to wait until the middle or the end of 2015 to proudly say that you use one.

NEC SX-ACE: 16 Tflops per vector cabinet

Is the vector supercomputer dead? Certainly not if one can judge by NEC's new SX-ACE, successor to the valiant SX-9, which increases computing power by a factor of 10 and reduces floor surface by 5. The machine features a proprietary vector SoC containing four 64 GB bandwidth vector cores delivering up to 64 Gflops each, which makes them simply the best-performing cores in the world! On the form-factor front, an SX-ACE cabinet can house up to 64 modules each containing one quad-core processor, for a total of 16 Tflops per cabinet. It is possible to interconnect up to eight, the maximum peak performance then reaching 131 Tflops. If, from a technical standpoint, SX-ACE represents a certain state of the art in massive parallelization, NEC now has to convince customers to exit the familiar x86 universe and adopt "new" programming methods using NEC compilers and libraries...

K40, CUDA 6: NVIDIA FOCUSES ON PERFORMANCE

It was SC 2013's big announcement for the GPU acceleration specialist. Thanks to the continuous improvement of manufacturing processes, the Tesla K40 replaces the Tesla K20X at the top of the Kepler architecture with double the memory (12 GB of GDDR5), a number of active cores brought to 2880 (instead of 2688), a slightly increased frequency (745 MHz instead of 735) and a doubled PCIe gen3 connection - all in the same 235 W TDP. As a result, official peak SP / DP performance now reaches 4.29 / 1.43 Tflops. According to various benchmarks presented by the manufacturer, the measurable speedups lie between 20 and 40% - the highest results being obtained on the most memory-hungry applications (CFD, seismic...). NVIDIA also announced a new, developer-controllable "GPU boost" mode designed to temporarily increase the frequency of all cores simultaneously (up to 875 MHz) should the application really need it.

Engagement before the sacred union? The new version of CUDA, also announced at SC, unifies memory logically, to the delight of developers who will no longer have to explicitly manage data transfers between the two distinct CPU and GPU memories. Nothing to do, therefore, with the physical unification of both spaces, which would definitively remove ze bottleneck, but this Nirvana remains on the roadmaps.

CUDA 6 is in da place

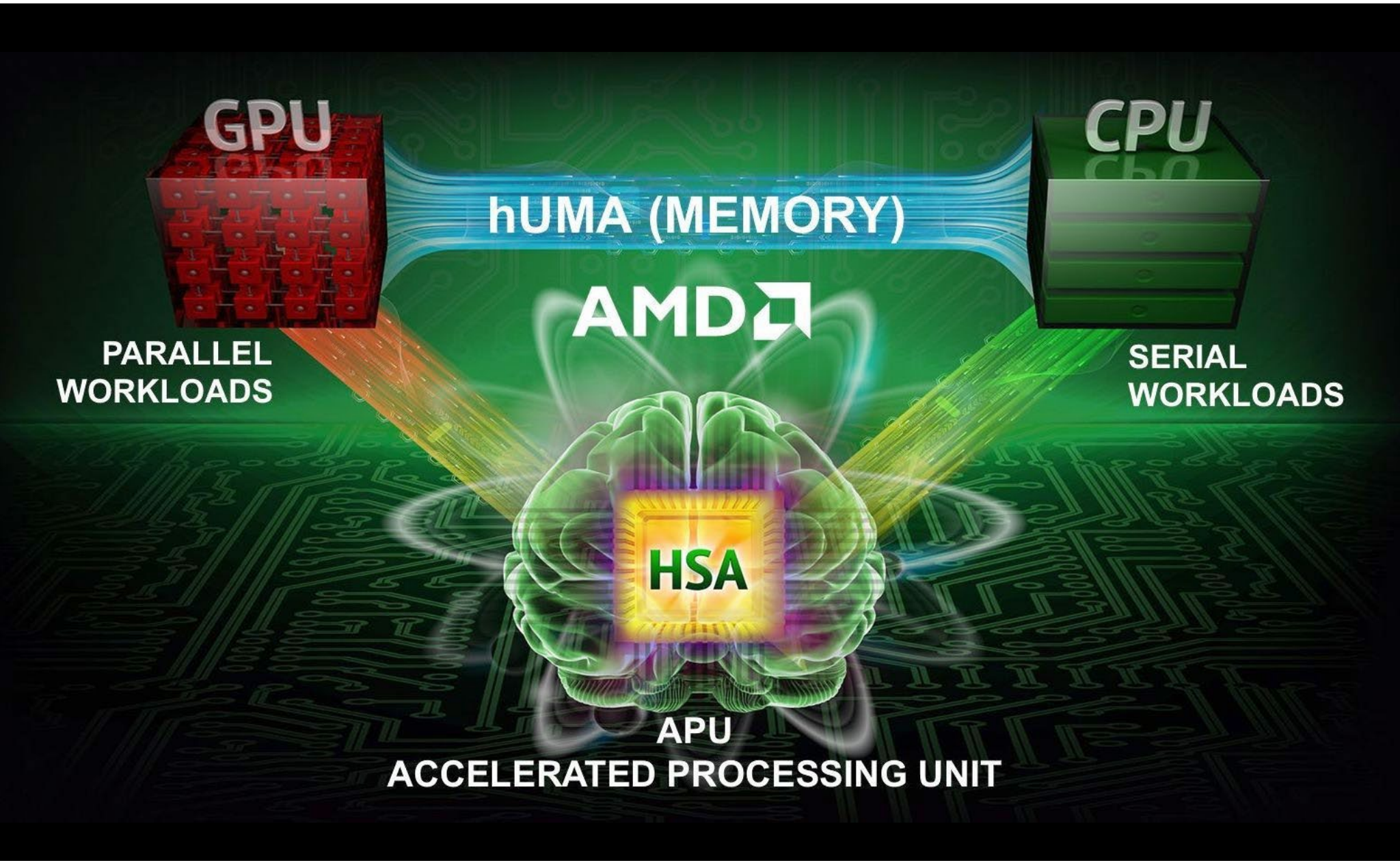
Beyond Unified Memory, CUDA 6 also provides drop-in libraries (automatic replacement of the CPU versions of BLAS and FFTW with their GPU equivalents) as well as the multi-GPU optimization of these libraries (up to 8 GPUs per node) and the support of larger workloads (up to 512 GB). An important release in other words, to which we'll devote an in-depth Programming feature (together with the HPC Labs) very soon.

Amazon Web Services adopts GPUs

AWS, the undisputed HPC as a Service world leader, expands its portfolio of remote computing offers with new Amazon EC2 (Elastic Compute Cloud) instances: the NVIDIA G2 ("NVIDIA GRID GPU") series. Executed through NVIDIA's Kepler graphic processors, these instances come in addition to the thirty or so services already offered to businesses and organizations in need of global or specific on-demand computing and visualization resources. NVIDIA G2s featuring a powerful H.264 engine dedicated to the compression of high resolution graphics, users should be able to work remotely on virtually any type of terminal, including mobile, regardless of the host application. Such a move by a top player like AWS should help demonstrate the practicality of "eHPC", convince conservative HPC users and therefore accelerate its adoption. At least within the United States, where the issue of homeland data hosting is not the same as in the rest of the world...

/news

AMD retakes the initiative



Already present in the desktop PC and game consoles, the AMD APUs (which combine CPU and GPU on the same die) are entering servers kingdom. With the Kaveri architecture, available in early 2014, some gains in density and energy efficiency are to be reasonably expected. To prepare the applications for this new technical deal, AMD and partners have also worked on the programming tools. Among new, adapted versions announced are the Java Virtual Machine, OpenMP compilers by GCC and PGI, OpenCL libraries like AMD's clMath and AccelerEyes' ArrayFire, as well as CodeXL, AMD's own debugging and profiling environment.

Beyond these announcements, AMD's vision is to unify all industry hardware platforms around HSA (Heterogeneous System Architecture), on which the next APU generations will be based. Bringing together other heavy-weights such as ARM, Ti, Qualcomm and Samsung, the **HSA Foundation**'s mission is to offer a single hardware abstraction for all heterogeneity needs. That should significantly ease and speed up the work of the developers: a single architecture is nothing less than the promise of a single executable code in a wide variety of environments...

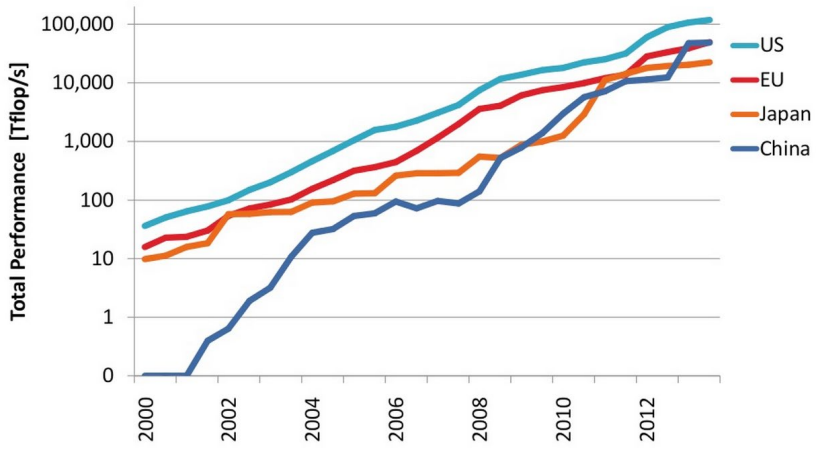
To achieve this unification effort, long-awaited by many in the community, a new OpenCL-based SDK was also announced during the last **AMD Developer Summit**. The SDK is derived from the APP SDK 2.9 and will include multimedia tools for video processing, the clMath library and CodeXL. This confirms, more than ever, AMD's commitment to Open Source for the programming of its processors and now those by others - an approach to high performance computing advancement through real portability that has probably been decried too quickly.

Nothing really new in the new Top500...

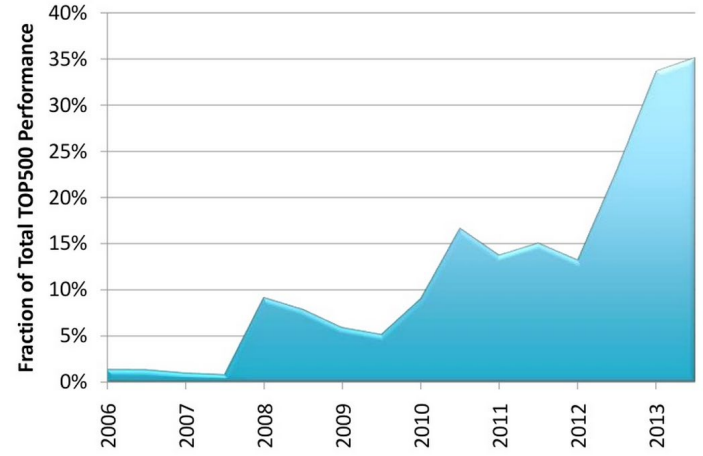
Intangible fatality, competition without excitement? The first among global supercomputers this semester is China's now well-known Tianhe-2 (34 Pflops) again, far - very far - ahead of Titan (18 Pflops), still ORNL's flagship for now. We noted the surprise entry within the top 10 (#6 to be exact) of Piz Daint, the Cray XC30 system of Switzerland's CSCS, with a remarkable energy efficiency of 2.6 Gflops / W for its total 5.6 Pflops peak performance. We also noticed that the list now counts 31 petaflops machines (+5), and that the number of clusters using hybrid technologies, 53, has not changed since June: 38 use NVIDIA Tesla GPUs, 13 include the Intel MIC (Xeon Phi) architecture and 2 rely on AMD accelerators.

From a geographical standpoint, nothing really new either. The United States is still largely dominating the global HPC landscape with more than half of the listed systems, followed by Asia (115 systems) and Europe (102). In Asia, China alone accounts for 63 entries, two less than in June. In Europe, the United Kingdom, France and Germany claim respectively 23, 22 and 20 of the most powerful systems.

Performance of Countries



Performance Share of Accelerators



November 2013's Top500 is roughly the same as June's. Even the growth rate of accelerator use is plateauing.

...but the Green500 gets greener

This fall's Green500, on the other hand, is completely renewed - to the point that it could be seen as a showcase for NVIDIA's accelerators (or at least a proof of their intrinsic greenness). Six months ago, there were only two Tesla GPU-based clusters in the top 10. Now, there are ten. For comparison, the first twenty machines in the June 2012 listing were IBM BlueGene/Qs. The most energy efficient system to date is called Tsubame-KFC (nothing to do with fried chicken). Installed at the Tokyo Institute of Technology, it reaches 4.5 Gflops per Watt. Following it is Wilkes (University of Cambridge), Europe's greenest computer with 3.6 Gflops / W. But the interesting point about the new list is definitely the global increase in power among "eco-friendly" supers. Whereas previous rankings showed relatively modest levels of peak performance, the top 10 now includes three petaflops machines. Things are progressing - certainly not as quickly as we'd like - but they are progressing...

/news

Jack Dongarra officially excellent (again)

A scientist well-known to the community for his work on software tools and computation libraries, Professor Jack Dongarra just received the Ken Kennedy Award for his "leadership in the design and promotion of standards for mathematical software used in solving numerical problems in high performance computing".

Created in 2009, the Ken Kennedy Award is a tribute to the late Ken Kennedy, founding Chairman of the computer science department at Rice University and a world-renowned HPC expert. The distinction, sponsored by the ACM and the Computer Society, goes to the scientists who contributed the most outstandingly to the programmability and productivity of high performance computing. This is exactly Jack's profile, whose name is associated not only with the creation of Linpack, the benchmark that has long been the reference for evaluating the performance of HPC machines, but also with the development and implementation of numerous world-class Open Source libraries, among which BLAS, LAPACK, Netlib, MPI, OpenMP-MPI, PAPI, ATLAS, PLASMA or MAGMA...

What is unique to Jack Dongarra is that his countless professorial duties have not prevented him from travelling the world, year in, year out, to generously share his personal insights with all who seek them. Animated by a unique blend of curiosity and creativity, he now works on a number of software issues connected with Exascale, and specifically with **PaRSEC** (Parallel Runtime Scheduling and Execution Controller), a scheduler for heterogeneous manycore architectures with distributed memories. We would like to humbly add our congratulations to those of the prestigious Jury.

Micron makes memory a parallel processor

Von Neumann may well be rolling over in his grave! Instead of constantly improving the bandwidth of the memories that feed our insatiable processors, why not reverse the problem and use memory not only as a simple space storage but also as an active processing provider? This is what Micron, who also invented the Hybrid Memory Cube, did with the Automata Processor. This literally revolutionary concept has the potential, at least on paper, to compete with coprocessors and accelerators in massively parallel computing.

What does it look like? Imagine an FPGA containing a grid of hundreds or even thousands of programmable computing units which, unlike a GPU running the same operation in parallel on Gigabytes of data, would be able to discretely parallelize different operations. The programming of such a component, i.e. the transformation of a given application into a dedicated machine, will not, however, be child's play. To facilitate it, Micron is currently working on a dedicated SDK that should be available before the summer. We'll keep following the story as it develops. Stay tuned!

The Exascale à la Mont-Blanc, Redux

SuperComputing 2013 has been the occasion for BSC's research group in heterogeneous architectures, led by Dr Alex Ramirez, to unveil a third prototype of Mont-Blanc exascale design associating power and energy efficiency. What do we find at the core of the machine? Low-power CPUs and GPUs, unsurprisingly of the very same kind as those that equip our mobile devices. This time, BSC chose to use Samsung Exynos APUs combining a dual-core ARM Cortex-A15 CPU to a Mali-T604 GPU on the same die. From an integration standpoint, This is a significant step forward in comparison to Pedraforca, the previous prototype unveiled last year, which contained twice less powerful quad-core Cortex-A9 CPUs (NVIDIA Tegra 3's) and used a separate NVIDIA Kepler K20 accelerator for parallel computations.

The new cluster also has the particularity of being designed around the Bullx B505 rack blade system and the corresponding new energy efficient chassis system developed by Bull. Each of the six clustered chassis contains nine blades which in turn house fifteen Mont-Blanc calculation cards. Globally, the 1620 ARM cores + 810 Mali GPUs offer a peak performance of 26 Tflops for 50% less electricity consumption than Pedraforca, a single blade delivering in excess of 485 Gflops for 200 Watts. But it is probably with the next iteration, which will feature the already available 64-bit ARM v8 CPU, that this design's energy efficiency will take on its full meaning. Thanks to a new 8.1 M€ funding envelope from the European Commission, the project is assured to continue for at least two additional years.

Linpack on its way to the shelf

Is the debate concerning the usefulness of Linpack about to be closed - once and for all? Developed at the end of the 1970s to measure floating-point performance, Linpack is clearly no longer representative of the behavior of our applications on contemporary machines. Hence the interest generated by Jack Dongarra, the undisputed authority in the domain since he created the benchmark, when he announced during SC'13 that he is working with his team on a new test protocol called HPCG (High Performance Conjugate Gradient). The goal, obviously, is to more accurately reflect the simulations performed today on large clusters. A first alpha version of the new benchmark will soon be available for validation on a broader spectrum of machines.

Dongarra warns, however, that although this alpha version has already produced usable figures, the time has not yet come to replace Linpack. A few quarters of running and fine-tuning will still be needed to make sure that HPCG acquires the general recognition of the community and thereby displace its ailing ancestor.

Nevertheless, the first HPCG results reveal spectacular performance gaps (sometimes by a factor of 40 to 50) compared with Linpack! Clearly, there is change underway in the routine Top500 ranking. The opportunity for a certain number of decision makers to focus more on actual users needs?

Supermicro: new record densities!

As NVIDIA officially announced its new K40, Supermicro introduced a new line of servers, blades and workstations with record levels of density: up to 20 accelerators in a 7U chassis!

Among the numerous references in the manufacturer's catalogue, we remarked the SuperServer which can house no less than eight GPUs and two Intel Xeon E5-2600 v2 Ivy Bridge CPUs, all in 4U. Given the necessary heat dissipation on such a design, two separate cooling systems have been integrated. Supermicro also offers a wide range of SuperBlade servers which, in 1U, 2U or 3U, will support a total of three to six GPUs.

As for workstations, the line is enriched by the SuperWorkstation - a racing machine with up to five integrated GPUs and redundant 1620 W Platinum power supplies that will likely be all the rage in Alaska and Greenland. Depending on the user's needs, it can also accommodate two Xeon E5-2600 (v1 or v2), up to 1 TB of DDR3 ECC memory, eight 3.5" storage units and an array of expansion cards (4 PCIe 3.0 x16, 2 PCIe 3.0 x8 and 1 PCIe 2.0 x4).

A new Allinea tool to analyze parallel performance

Well-known in the community for its DDT debugger and MAP profiler, Allinea took advantage of SuperComputing '13 to unveil Allinea Performance Reports, a new execution performance analysis environment tailored to parallel HPC applications. Its main objective: to determine if the codes are adapted to the hardware platforms on which they are meant to run, and to help developers identify bottlenecks and other weak spots. The good idea, here, is that the platform does not require the code to be instrumented or recompiled.

For good measure, Allinea also announced the support of new targets for its suite, including Intel's Xeon Phi and ARM's v7 architecture through the NVIDIA CUDA 5.5 platform. Logically, CUDA 6 should follow shortly...

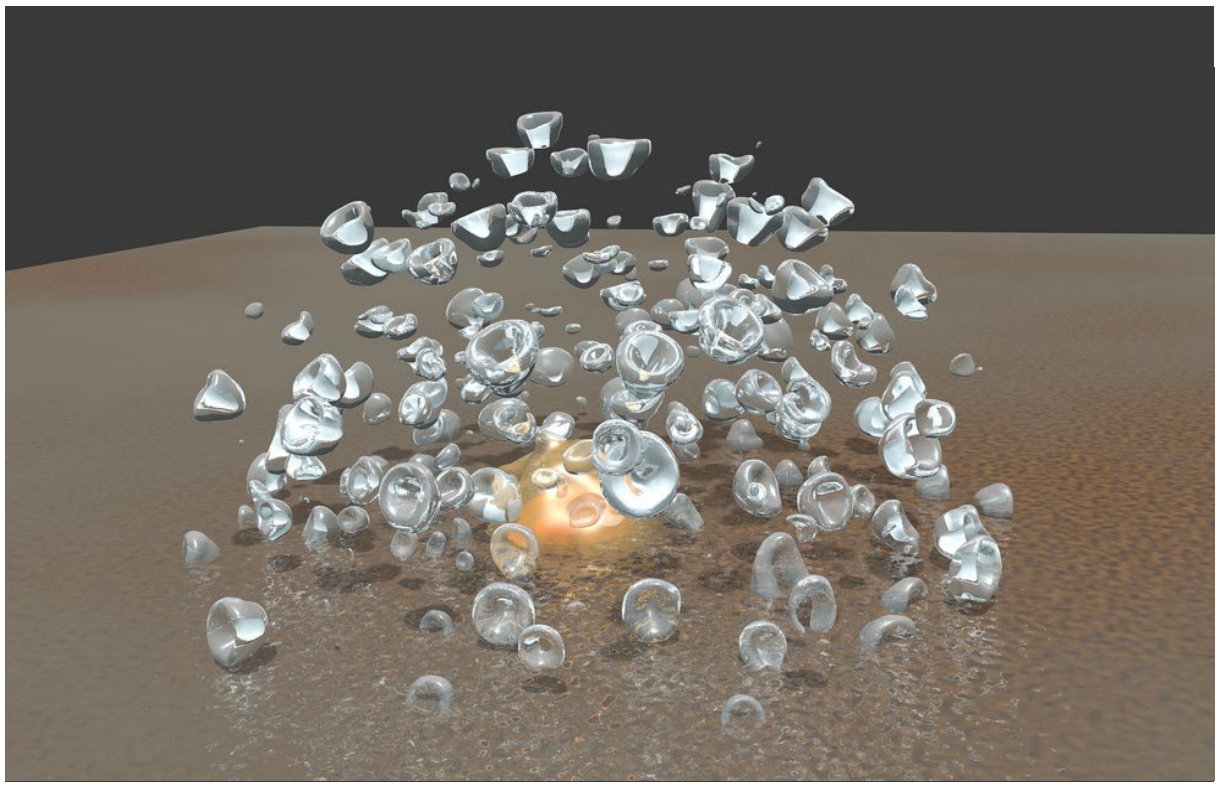
Rent-a-CAPS!

CAPS Entreprise, the HPC compiler technology specialist, just announced the creation of its new CAPS Engineering entity. Along with parallel computing courses in CUDA, OpenCL, OpenMP, MPI and processor parallelism, CAPS Engineering will provide custom services such as the diagnosis of application parallelism and the porting and fine-tuning of codes to parallel systems with or without accelerators, and domain-specific customizations. "*In view of the different paths the parallel processors are taking, we see a growing demand in parallel programming expertise*" said Yann Mevel, General Manager at CAPS. All the best!

/news

A few Bubbles for the Gordon Bell Prize

Of the six finalists of the 2013 Gordon Bell Prize, four used Titan to realize their simulations. But statistics weren't consulted for the attribution: the distinction was awarded to a team of scientists at ETH Zurich and IBM Research, in collaboration with the Technical University of Munich and the Lawrence Livermore National Laboratory (LLNL), for their work on cavitation carried out on Sequoia, LLNL's IBM BlueGene/Q.



The experiment conducted by Petros Koumoutsakos and his colleagues resolved unique phenomena associated with clouds of collapsing bubbles. This condition occurs when vapor bubbles formed in a liquid collapse due to changes in pressure. The successful effort employed 13 trillion cells and 6.4 million threads, resulting in a 14.4 Pflops simulation (approx. 73% of the machine's peak power) that resolved 15,000 bubbles and achieved a 20-fold reduction in time to solution over previous research. The results will help improve the design of combustion engines and hydraulic turbines. They should also positively impact a number of therapeutic approaches to cancer. Kudos to the team, and also to the other finalists whose selection will certainly promote their career.

Who wants to win an AMD FirePro S10000* 6GB accelerator?

(Get lucky, there's 2 to be given away this month)

Fits comfortably in most OpenCL developers workstations.

Two high-end AMD GPUs directly on-board for powerful dual GPUs programming.

Free AMD Accelerated Parallel Processing SDK, OpenCL 1.2 compliant with BLAS and FFT libraries.

OpenCL 2.0 ready!

New! Free AMD CodeXL 1.3 with CPU-GPU debugger, profiler and static OpenCL code analyzer.

Future-minded developers want a parallel computing architecture that does not restrict their ability to use open tools and APIs to develop cross-platform.

Managed by The Khronos Group, like the widely used OpenGL framework, OpenCL is the solution to this legitimate demand. OpenCL is all about easy code portability, which makes it the only future-proof path to address all HPC coding requirements.

Win yours @ www.hpcmagazine.com

*AMD FirePro S10000 6GB Active technical specifications :
5.91 TFlops peak single precision, 1.48 TFlops peak double precision, 6GB GDDR5 memory, 480 GB/s memory bandwidth, Dual-slots form factor, to be used in a workstation.

AMD FIREPRO

Improving efficiency is the smarter way to operate.

On a smarter planet, businesses must accelerate growth and improve profitability. Through a unique combination of industry experience and expertise, IBM is helping manufacturing companies address critical issues across the value chain. With a strong global presence, advanced research and development capabilities and comprehensive hardware, software and services, we are equipped to help you improve operational efficiency as well as health, safety and environmental practices, manage costs and become more customer centric. IBM has the tools, technology and people to help you meet today's manufacturing challenges.

A smarter business needs smarter thinking.
Let's build a Smarter Planet.

View this executive summary of the results of a new survey conducted by Desktop Engineering to gauge audience familiarity with high performance cluster computing and its benefits. **http://bit.ly/14zq83c**

ibm.com/platformcomputing

/viewpoints/conway

TEN THINGS CIOs SHOULD KNOW ABOUT HPC

The reluctance and misunderstanding of IT departments with regard to HPC is a major obstacle to its corporate mass adoption. Here are ten arguments that should lead your CIO to enlightenment...

For organizations that already exploit HPC, HPC budgets are often at the mercy of chief information officers (CIOs) or other senior management officials who may not fully understand the nature and value of high performance computing. The same holds true for executives of organizations that haven't adopted HPC yet. With that in mind, here are ten things that these executives should consider.

1 - HPC Is One of the Fastest-Growing IT Markets

Between 1990 and 2012, revenue for the worldwide HPC ecosystem - servers, storage, software, and services - ballooned more than tenfold, from $2 billion to $21.9 billion. Propelled by the standards-based cluster - an HPC innovation from NASA - the HPC server market expanded during the decade of the 2000s faster than the "hot" IT markets for flat-panel TVs or online gaming. The worldwide HPC server market posted record revenue in 2011 and 2012, with server systems costing $500,000 and up, achieving a 29.6% year-over-year revenue gain.

2 - Commercial Firms Began Adopting HPC in the 1970s

Following its initial growth spurt in the late 1960s and 1970s, the market for HPC systems has expanded over time by adapting to the requirements of successive waves of new users - in large part through advances in software. In 1976, the CRAY-1 supercomputer was delivered to its first customer, Los Alamos National Laboratory (LANL), as a blazingly fast hardware platform with no operating system. The second wave of adoption carried HPC into industry, initially the automotive, aerospace, and petroleum sectors, starting in the late 1970s. The third adoption wave, driven by the compelling price/performance of standards-based clusters, began in 2001-2002 and continues today. This phase greatly expanded the market for HPC by making this game-changing technology affordable and tractable for less-experienced users, such as financial services firms, consumer products makers, and online companies of nearly any size.

3 - 97% of Adopters Say HPC Is Indispensable for Their Ability to Compete and Survive

In a worldwide IDC study conducted for the Council on Competitiveness, based in Washington, D.C., 97% of the commercial firms that had adopted HPC said they could no longer compete or survive without it. The chief benefit cited by these firms is that HPC enables them to bring more innovative, higher-quality products and services to the market in shorter time frames. To out-compute is to out-compete.

4 - Senior Government Officials Increasingly Recognize HPC's Economic Value

Worldwide political leaders increasingly recognize this trend:

• In his 2006 State of the Union address, U.S. President George W. Bush promised to trim the federal budget, yet urged more money for supercomputing.

• In 2009, Russian President Dmitry Medvedev warned that without more investment in supercomputer technology, Russian products "*will not be competitive or of interest to potential buyers.*"

Think of HPC not as an IT function but as a competitive business advantage. There's a hard link between HPC and a company's top line and bottom line.

• In February 2012, the European Commission announced that it had adopted a plan to double spending on HPC to €1.2 billion, with much of that money aimed at the installation of additional large supercomputers at leading European HPC centers.

5 - HPC Is Different from Business Computing

HPC jobs are different from business operations such as accounting, payroll, sales, customer relationship management (CRM), enterprise resource planning (ERP), transaction processing, human resources, and purchasing. The common denominator underlying HPC problems is a degree of algorithmic complexity that is atypical for business IT problems. Business IT workloads often consist of a large volume of tiny problems - for example, a business computer may process tens of thousands of transactions per second. In contrast, a single HPC problem may take not a fraction of one second but hours, days, weeks, or even months to process.

6 - The Goals of IT and HPC Are Also Different

CIOs who are new to HPC often make the mistake of treating it like a typical IT function. This misperception can lead to lost productivity and to conflict between the CIO's office and the company's HPC staff.

I recently interviewed Jim Barrese, CTO of eBay's PayPal subsidiary, which adopted HPC not long ago for real-time detection of online consumer fraud. In that interview, he advises CIOs to "*clearly understand that HPC is not a mass consumption technology where we enable everyone in our organization with it. This is a deep engineering function. It's custom built and includes writing software to solve cutting-edge problems... Think of HPC not as an IT function but as a competitive business advantage. There's a hard link between HPC and a company's top line and bottom line.*"

In important respects, HPC is different from general IT deployments. IT is generally about provisioning - equipping each of the company's knowledge workers with the basic computing tools they need to perform their jobs productively, and providing as little beyond that as possible to stay within the budget. HPC, on the other hand, is about enablement - providing a small subset of specialized knowledge workers with the most powerful computational tools the company can afford. A typical IT worker's PC is capable of fully supporting the worker's computing requirements, while there is often no limit to the amount of computing power an HPC user could exploit on the company's behalf.

7 - Key IT Datacenter Technologies Have Trickled Down from HPC

There is a perennial debate between those who argue that key IT technologies "bubble up" from the low end, such as embedded and desktop devices, and those who counter that key technologies "trickle down" from the high end, especially HPC. In reality, of course, both arguments are correct. Technological innovation is bidirectional, flowing up and down. For example:

• Standard x86 processors bubbled up into HPC from the market for desktop/laptop computers.

• Conversely, x86-based clusters were born in the HPC market and later trickled down into enterprise IT datacenters.

With entry prices below $10,000, HPC systems have become affordable for many more companies, small and medium, than ever before.

• The Linux operating system is an HPC innovation that helped make clusters dominant in HPC. Linux clusters later began moving into financial services firms and other commercial datacenters.

• Grid computing and cloud computing are two more important technologies that have trickled down from HPC to mainstream commercial markets.

• On the bubble-up front, multiple processors and coprocessors have been making their way from the embedded systems market into HPC, including GPUs, ARM, and Atom devices.

8 - HPC Systems Now Start at Under $10,000

Decades ago, entry pricing for a supercomputer was in the $25 million to $30 million range. Thanks to the transition to clusters based on industry-standard technologies, pricing for HPC systems now starts at less than $10,000. With entry prices this low, HPC systems have become affordable for many more companies than ever before.

9 - Commercial Firms Are Also Adopting HPC for Challenging Big Data Problems

High-performance data analysis (HPDA) is the term IDC coined to describe the convergence of the established data-intensive HPC market and the high-end commercial analytics market that is starting to move up to HPC resources. The financial industry has been running analytics on HPC systems at least since the late 1980s. But newer methods, from MapReduce/Hadoop to graph analytics, have greatly expanded the opportunities for HPC-based analytics.

A 2013 IDC worldwide study showed that 67% of HPC sites are running HPDA workloads. IDC forecasts that revenue for HPC servers acquired primarily for HPDA use will grow from $739 million in 2012 to $1.4 billion in 2017. Revenue for the whole HPDA ecosystem, including servers, storage and interconnects, software, and service should double the server figure alone.

10 - There Is More on Tap from HPC

One of the next important developments IDC expects to come out of the HPC market is more capable network technologies to speed communications between cores, processors, servers, and nodes. This development should help to address the so-called memory wall, the growing gap between escalating processor peak speeds and the lagging ability of internal networks to feed processors with enough data to keep them busy. Improving network bandwidths and latencies should be especially important for challenging Big Data tasks faced by businesses and government organizations alike.

What Does This Mean?

As more companies of all sizes in more markets learn to exploit HPC (and its close relative, HPDA) to speed and improve innovation, competitors lacking this advantage will fall behind. Successful CIOs will therefore need to gain a basic understanding of HPC and ensure that their organizations carefully consider whether to adopt this technology.

Steve CONWAY
IDC Research Vice President, HPC

/viewpoints/gentzch

HPC AS A SERVICE?

To discover, explore, and understand the end-to-end process of accessing and using HPC Clouds - such is the aim of the UberCloud HPC Experiment. With more than 800 organizations from 66 countries now participating, the operation demonstrates the viability of this alternative to traditional and costly on-premises high performance computing.

As is well-known to this audience, the benefits for small and medium size enterprises (SMEs) of using HPC within their design and development processes can be huge: better quality products, high Return on Investment (ROI), reduced product failure early in design, shorten time to market... Potentially, this leads to increased competitiveness and innovation.

Why then are many engineers and scientists running simulations just on their workstations, although many are often dissatisfied with the performance? The main reason is that the alternatives are still coming with a lot of challenges for engineers and scientists.

Alternative one, buying an HPC server, comes with high TCO as has been regularly demonstrated by IDC: in addition to server cost, expenses for staffing, training, software, downtime, and maintenance easily sum up to ten times the server cost over three years. There are also long and painful internal procurement and approval processes. And for many, ROI is not clear, although it is expected to be huge according to countless recent studies.

The second alternative is recently offered by cloud computing. HPC in the Cloud (or HPC as a Service) allows engineers and scientists to continue using their own desktop system for daily design and development, and to submit (burst) the larger, more complex, time-consuming jobs into the Cloud. Benefits of the HPC Cloud solution (in addition to HPC in general) are among others on-demand access to virtually infinite resources, pay per use, reduced capital expenditure, greater business agility, not to mention dynamic resources scaling.

However, HPC as a Service comes with challenges of its own. It is a new business and working paradigm, for the manager as well as for the engineer. Security, privacy, and trust in service providers can be an issue. Conservative software licensing is only slowly including the pay-per-use service model. Internet bandwidth is often not able to accommodate heavy data transfer needs. Unpredictable costs of cloud computing can be a major problem in securing a budget for a given project. And finally, there is often a lack of easy, intuitive self-service access and use of remote resources.

Now, here comes the **UberCloud HPC Experiment** which provides a platform for engineers and researchers to discover, explore and understand the end-to-end process of accessing and using HPC as a Service, and to identify and resolve the roadblocks. Guided through a 22-step process, end-users, software providers, resource providers and HPC experts are collaborating in teams, jointly solving the end-user's application in the Cloud.

Since July 2012, the UberCloud HPC Experiment has attracted 800+ organizations from 66 countries, allowing to build 122 teams in CFD, FEM, Life Sciences and Big Data. Recently, the UberCloud University and a Virtual Exhibition have been added, and an Intel sponsored **Compendium** with 25 case studies has been published.

SMEs who are aiming to develop better products faster are invited to join the **free** UberCloud HPC Experiment.

Prof. Dr. Wolfgang GENTZSCH
Chairman, ISC, & Cofounder, The UberCloud

CARRI Systems

Technology architect since 1992

intel Technology Provider Platinum 2013

NVIDIA PARTNERFORCE TESLA PREFERRED PROVIDER

Conception Analysis Realization for Research and Industry

Founded in 1992 by enthusiasts, CARRI Systems is a French company recognized for its technological expertise in high performance computing systems. Based in Noisy-le-Sec near Paris, CARRI Systems is a member of the European Pole of Competence in high performance simulation (Ter@tec) since 2011.»

Franck Darmon

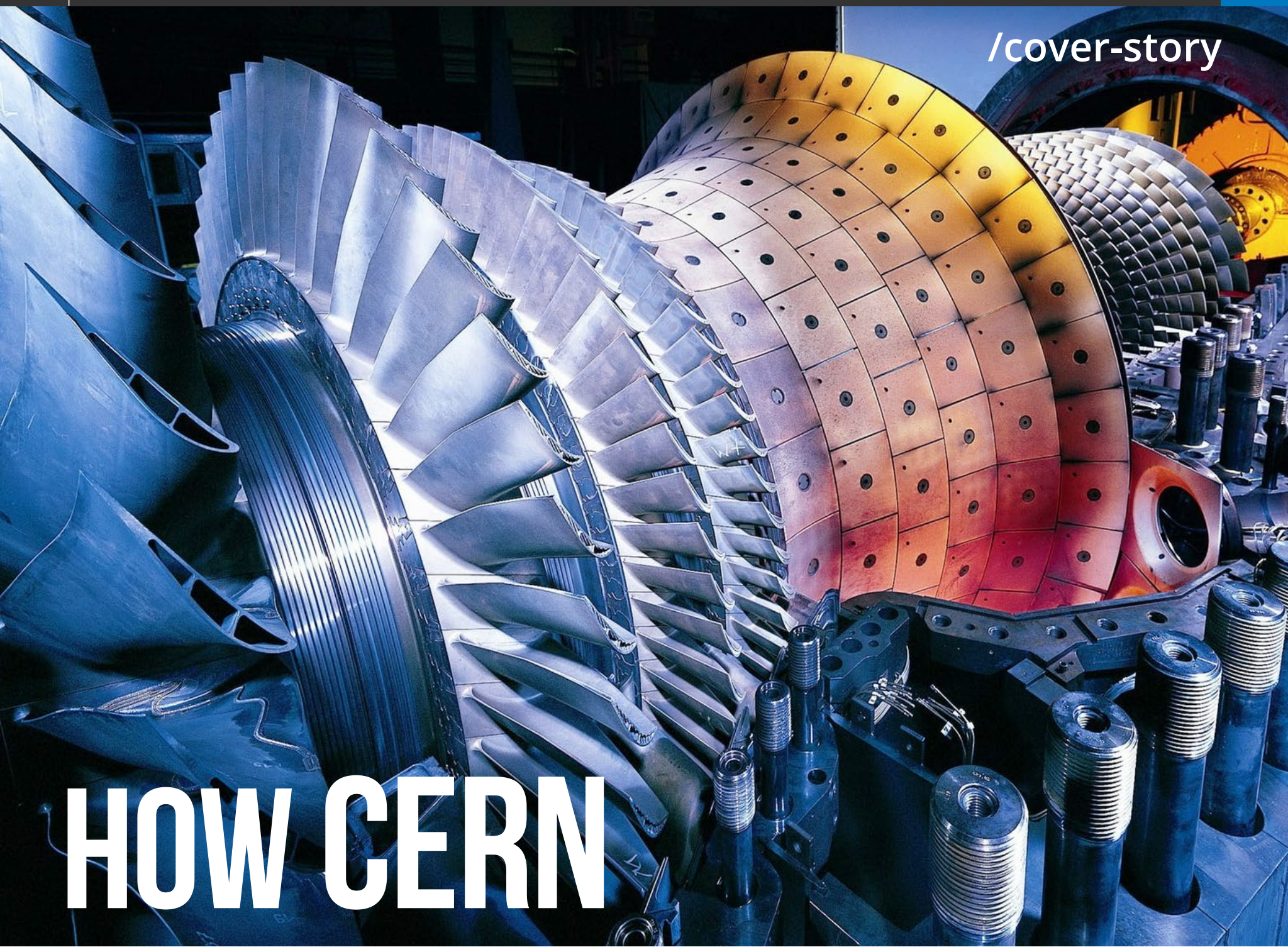
CEO, CARRI Systems

FABRIQUÉ EN FRANCE

intel NVIDIA AMD

For more information about XLR solutions

www.carri.com



/cover-story

HOW CERN MANAGES ITS DATA...

After 3 years of faithful service, after the more than probable discovery of the Higgs boson, the LHC starts its first long shutdown. The ideal occasion to discover how, on a day-to-day basis, CERN takes up the challenge of scientific Big Data...

FREDERIC MILLIOT

As you read this, more than a thousand researchers all over the world are working live on data from CERN'S LHC (Large Hadron Collider). If the Centre is universally recognized for its frontier work in fundamental physics, it is beginning to be recognized also for its compute infrastructure. With an annual usable production of experimental data exceeding 30 Petabytes, the problem is not simple, all the more as this wealth of knowledge must be shared without constraint of place or technology. The case of CERN being in many respects representative of the problems that academic and private organizations face managing large volumes of information, it seemed appropriate to inquire into the matter. It is to a guided tour of CERN's data installations that we invite you now.

Big Science > Big Data

"*The biggest challenge here in the CERN's IT department*", points out Frederic Hemmer, the engineer in charge, "*is obviously the*

volume of the data, particularly in production." The Centre must collect, analyze, store and protect this data to make it available 24/7 to community researchers. In this context, The choice of an agile infrastructure (that we will detail below) has proven relevant: to date, CERN has never lost a file, for any reason. And this although "*even when the LHC is not operating*", Frederic Hemmer emphasizes, "*IT does not stop. Our analyses are continuous, here at the Centre and all over the world.*"

Concretely, CERN must anticipate the very particular, and often rather inventive, needs of users who conduct experiments for which the technical requirement are generally unpredictable. This leads the IT teams to innovate day after day, with extreme result constraints and an operational budget that remains very... constant.

Oracle everywhere

Freely choosing the basic elements of its technological stack, CERN takes care to respect a theoretical balance between performance, reliability and scalability. For the data, the choice of an Oracle base was made as far back as 1982, and extended since then to all dimensions of the organization and operation of the Centre, including the systems controlling the accelerators during the experiments. According to the technicians responsible for these decisions, Oracle meets the prerequisites as regards functionality, availability and dimensioning. They appreciate that the software infrastructure includes the tools necessary for management, protection and distribution of the data.

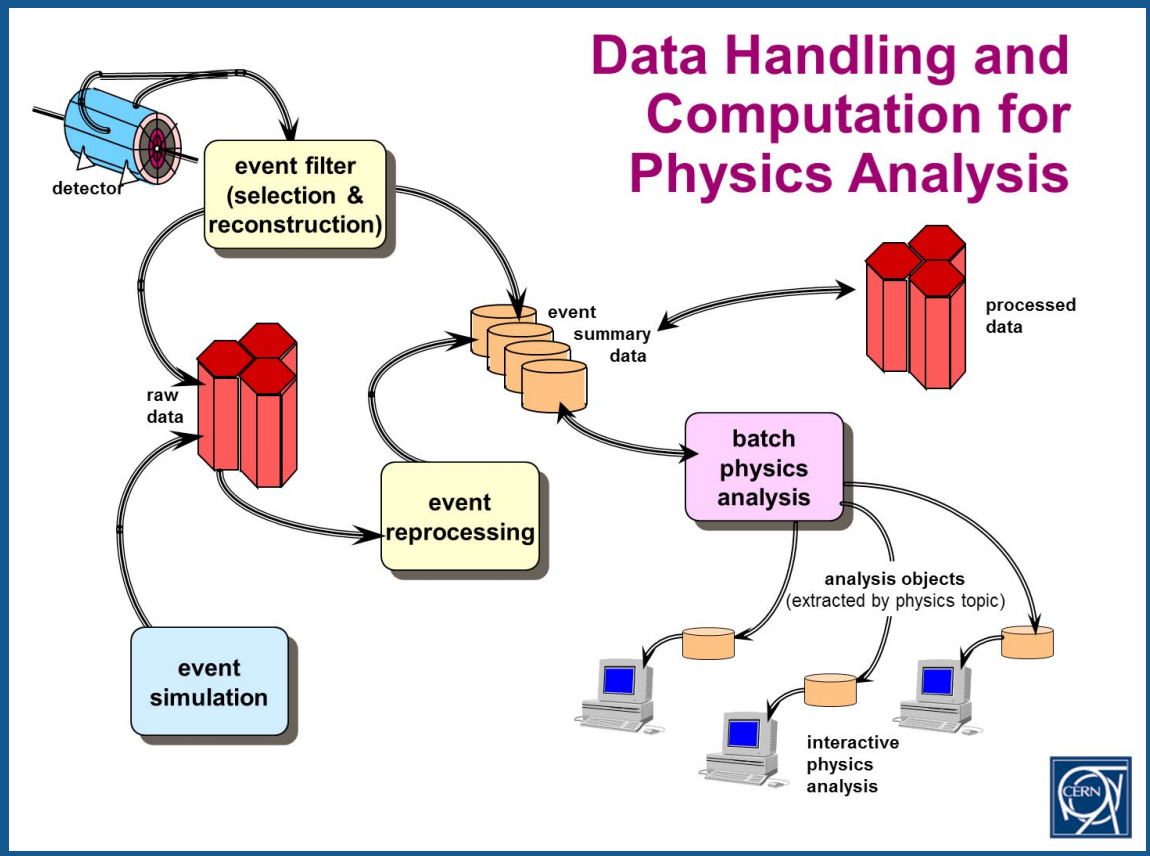
On the storage side, NetApp technologies constitute the bulk of the new facilities. For NetApp, such a showcase is

Data models for the LHC

Every LHC experiment uses a large number of sensors, so the number of interactive data structures is quite significant. Each sensor produces its own raw data, coming directly from the electronic measuring. The problem is that the analyses are typically much higher with regard to the abstraction levels than an elementary change of state. Therefore, complex reconstruction procedures must be executed, which in turn create new data.

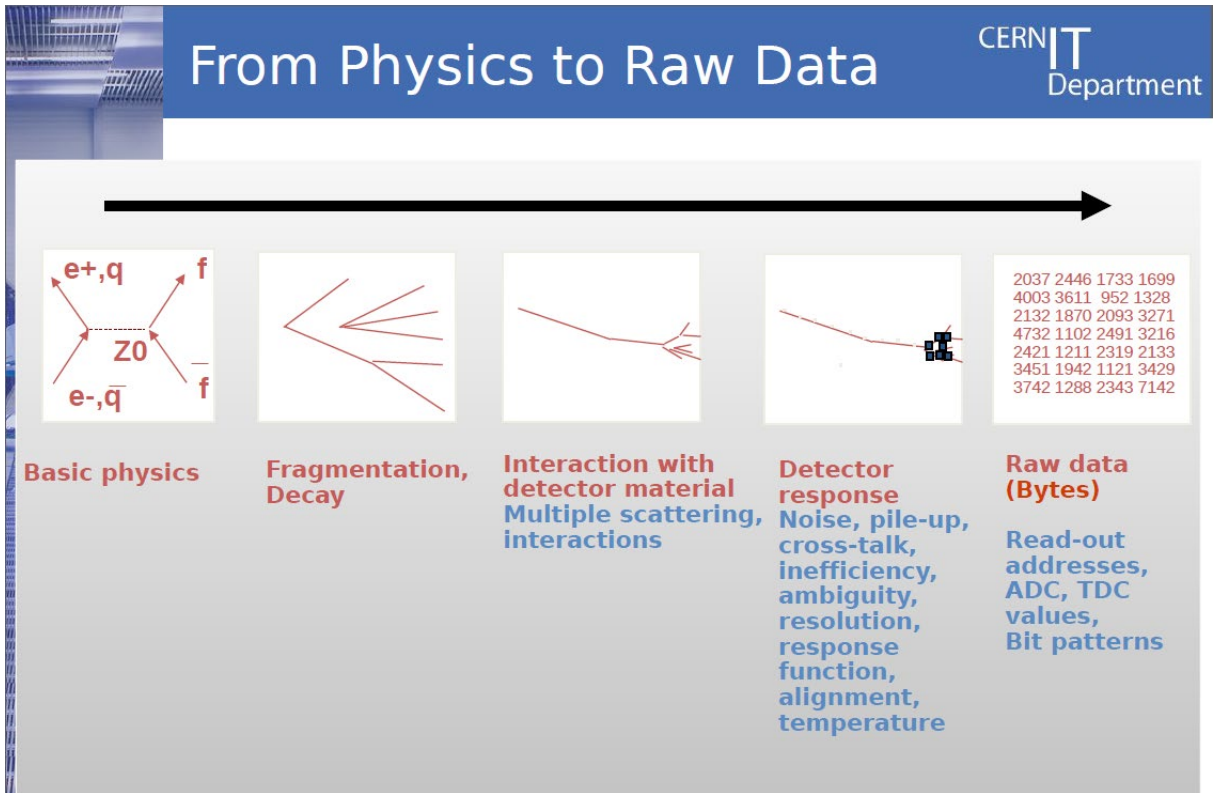
Considering that the events analyzed (such as the collision of particles, typically) are by principle independent, CERN records one data tree per collision. This tree contains data from all the sensors at different abstraction levels. In practical terms, this is equivalent to thousands of objects of discrete classes offering different types of relationships between them.

The analysis performed after the collection involves navigating through this network of objects from a wide variety of paths. The image below shows one of these paths (very simplified), which is to select a specific track from the list of all of the collision tracks, and then to follow all the status changes that made it up.



not without technical challenges. For example, collisions of heavy ions (lead nuclei, generally) are complex experiments that make information production rates very difficult to estimate in advance, but that can reach 6 GB per second.

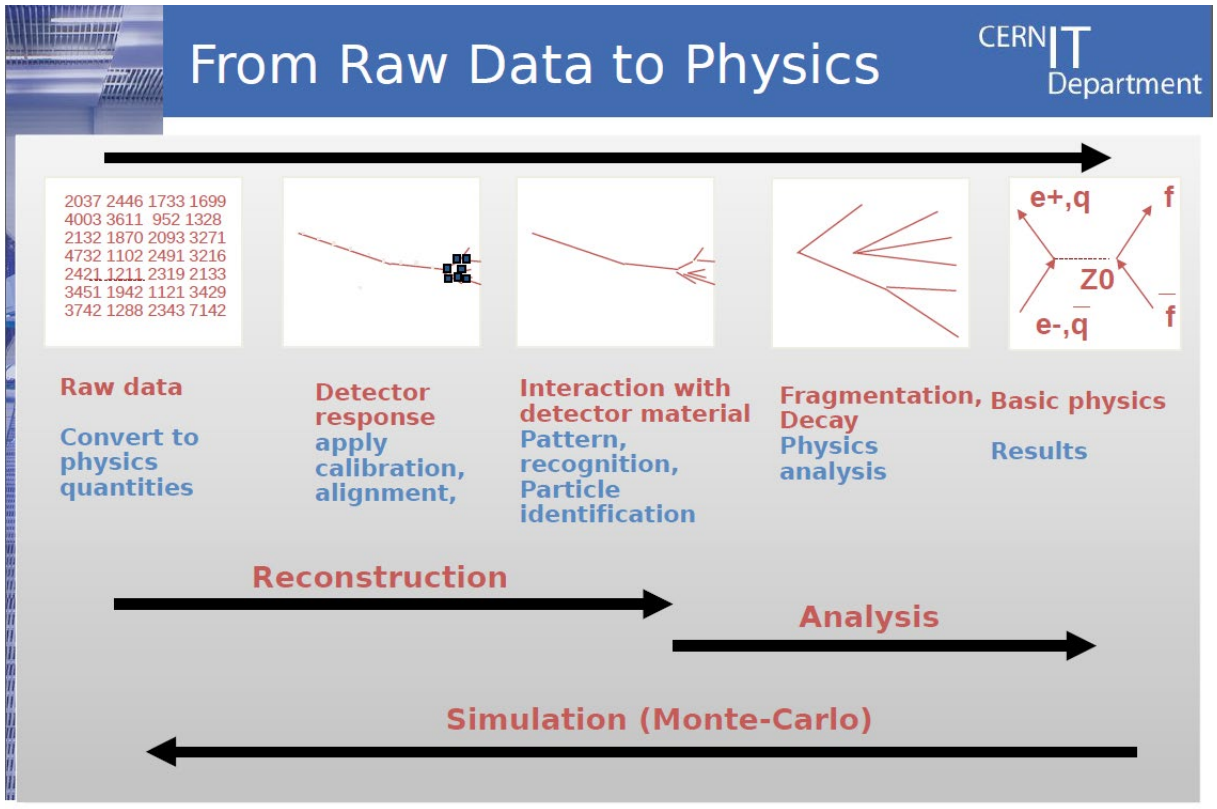
However, still according to Frederic Hemmer, "*data being the very reason for our existence, the mission of the IT teams is multi-dimensional. We must allow their almost immediate use, guarantee their immortality, manage equipment and software upgrades in a non-disruptive way and make sure that the infrastructure offers virtually infinite space and scale up capability.*"



The chain of collection for experimental data. In this direction, the operations primarily focus on the filtering by coherence, resolution and relevance. This filtering also has an important role to play with regards to the disambiguation of detected phenomena. At this stage, the throughput aspect takes precedence over the calculation aspect.

The LHC depends on its DBMS

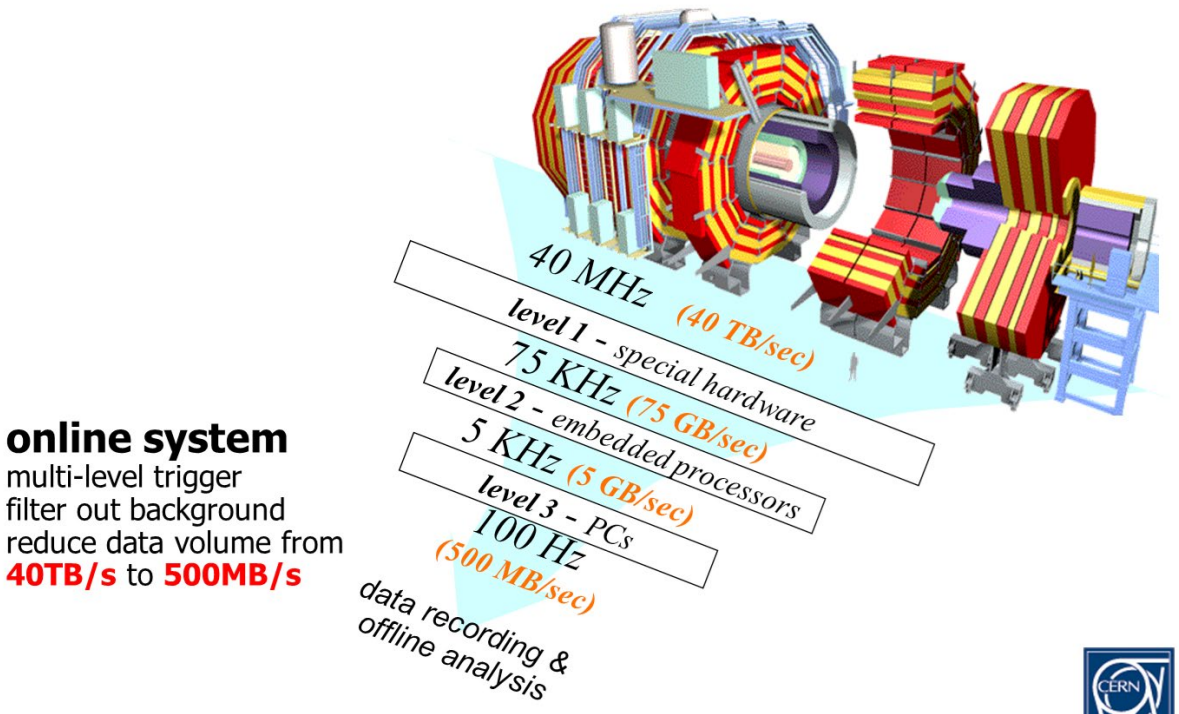
If such a statement is usually readily admitted, it should be well understood here that even the least problem in collecting or managing data implies stopping the collider. Technically, the collider is actually managed through two databases. The first, ACCCON, stores the setup and control elements of the installation. To ensure permanent monitoring of the LHC, the operators refine its configuration in real time through an array of control monitors grouped in a dedicated room. If the database is unavailable even for a few seconds, the collider becomes uncontrollable. It is therefore necessary to stop the experiment in progress, which implies killing the beams in enormous graphite blocks to disperse their energy and protect the ring. The extreme temperatures are, in fact, liable to damage the magnets (each of which costs more than a million dollars), which would require repairs likely to make the LHC operationally unavailable for months.



In the other direction, the reconstruction of phenomena from the data initially collected also supports simulation analysis. It is in this sense that data operations are the most demanding from a computation standpoint.

The second database, ACCLOG, is a registry of the inputs coming from the thousands of sensors which also make up the LHC. It is this database that contains the long-term logs of the state of the magnets and

the moving parts, particularly the collimators that protect the beams by eliminating scattered particles. With more than 4.3 trillion rows, this database is by far the largest (and the one that is growing the fastest) among all of the Centre's information systems. And of course, as it determines the calibration of the whole infrastructure, it is essential to maintaining the LHC online.



In the production line of experimental data, a multilevel filter system eliminates more than 99% of the information collected. Only the remaining information will be stored for analysis. CERN document.

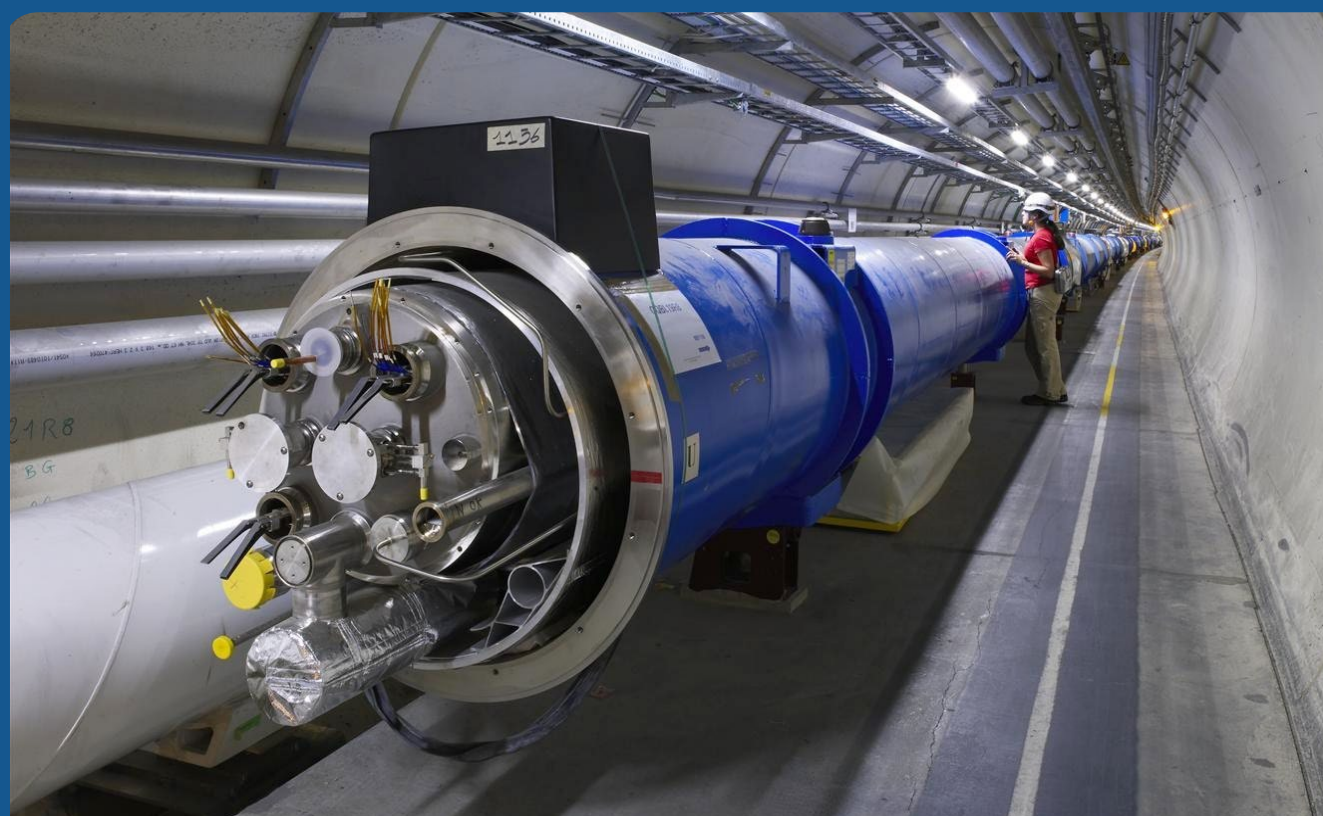
A needle in a thousand haystacks

Let us now put ourselves in the post-experimental phase. For the IT team, the mission is to offer the best possible performance in accessing the databases that index the raw data. This is recorded in ROOT flat files, which are then stored in a hierarchical system called CASTOR (see our two detailed boxes). The context, but also the processing, is thus almost

The LHC for dummies

Among the particle accelerators at CERN, the Large Hadron Collider (LHC) is certainly the most famous. Inaugurated in 2008, it forms a 27 km (16.13 miles) long circle adjacent to the Franco-Swiss border, about 100 meters underground the Alps.

The ring consists of superconducting magnets and acceleration structures designed to increase the particles' energy. Inside the ring, and then inside dedicated containers, the beams circulate in opposite directions. It is the magnetic fields from the superconducting magnets that guide them, the magnets being themselves cooled by liquid nitrogen and liquid helium so that their temperature remains at about -271°C. The infrastructure has been designed for the beams to collide in certain areas of the ring equipped with detectors. Several international teams are working on separate experiments involving specific detectors, which allows them to conduct collaborative studies on the material produced by the collisions. In cruise-mode, the LHC generates about 600 million collisions per second. Dedicated codes convert raw data into data objects for later analysis. Today, the annual volume of new exploitable information reaches in excess of 20 PB.



identical to those of a large Big Data application. More and more, to extract pertinent information from the huge mass of available data, datasets are run through starting from predictive analyses.

In practice, the system rests on five "critical" functional pillars, as Eric Grancher, the Database Architect in the IT team, explains:

• 10 GbE connections, a standard technology that is both scalable and easy to manage. These connections allow using the same switches as in many other subsystems in the Centre and offer the advantage of being maintained by CERN's network team, which provides 24/7 support.

• Plain SATA disks coupled to a special cache technology that offers a level of performance comparable to Fibre Channel drives at a fraction of their cost. Considering the number of disks used in CERN (more than 64,000, see the box below) the cost aspect is really determining here.

• The Oracle dNFS file system, which offers several storage paths. dNFS has the particular feature that it bypasses the operating system, which results in an almost double performance level, with another advantage of weight: Oracle generates NFS requests directly from the databases, so the system requires almost no configuration or maintenance.

• An instantaneous volume cloning technology that allows the creation of modifiable copies of datasets on the fly. This technology, called FlexClone, is of NetApp origin. The instantaneous aspect offers a second benefit in such a particular context: whatever the number of scientific teams working on a dataset, no data is duplicated.

Numbers unlike anything else:

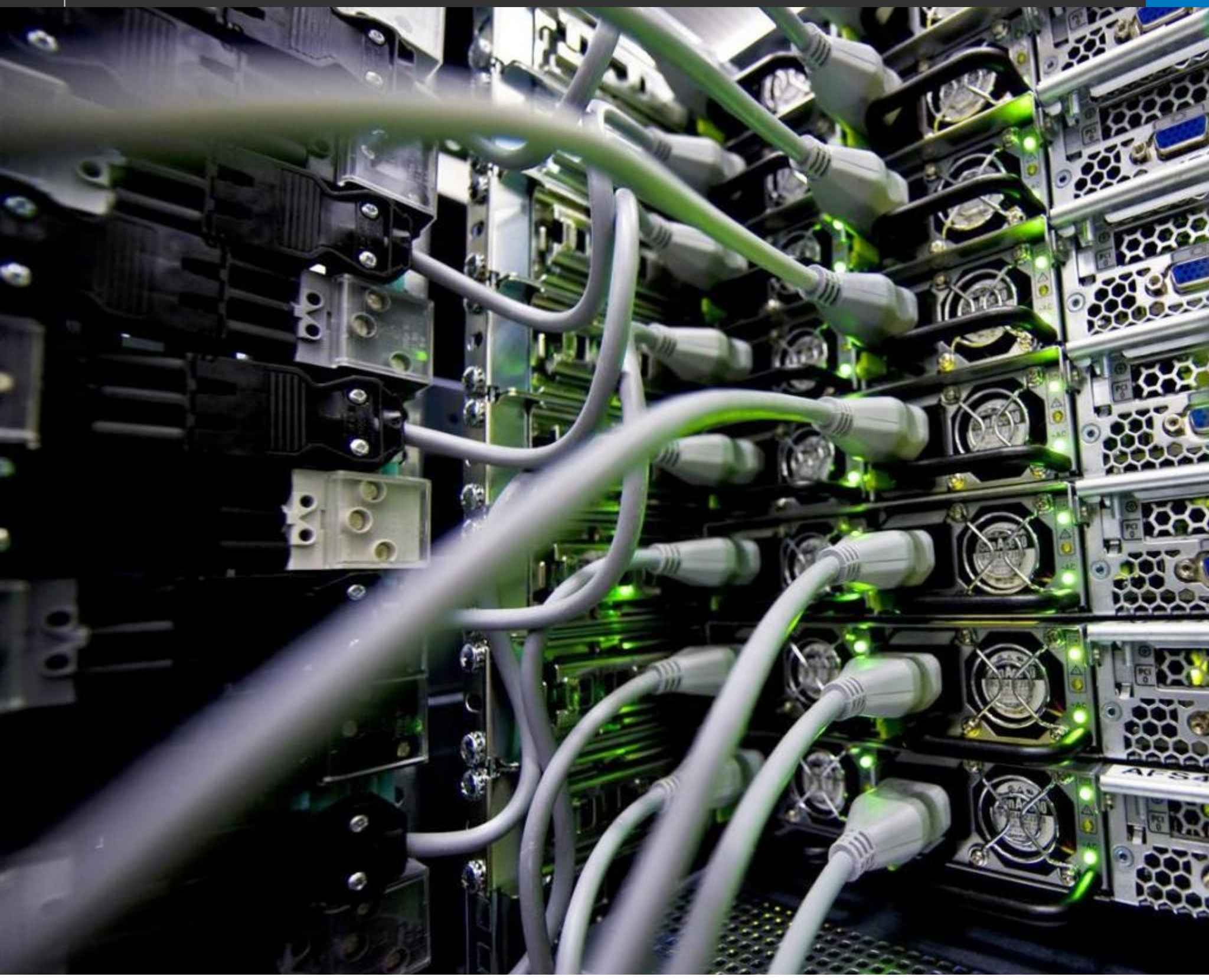
2,500 employees
10 000+ researchers and students
608 associate universities
113 nationalities on site

828 racks
11,728 servers
15,694 processors
64,238 cores
56,014 memory modules
158 TiB* of memory capacity
64,109 hard disks
63,289 TiB* of raw disk capacity
3,749 RAID controllers
1,800 disk failures / year
45,000 cartridges
56,000 cartridge slots
73,000 TiB* of tape capacity

24 high-speed routers (640 Mbps - 2,4 Tbps)
350 Ethernet switches
2 000 10 Gbps ports

2,4556 MW (IT use)
120 MW (LHC use)

* 1 TiB (tebibyte) = 2^{40} (1 099 511 627 776) bytes, or 1024 gibibytes.

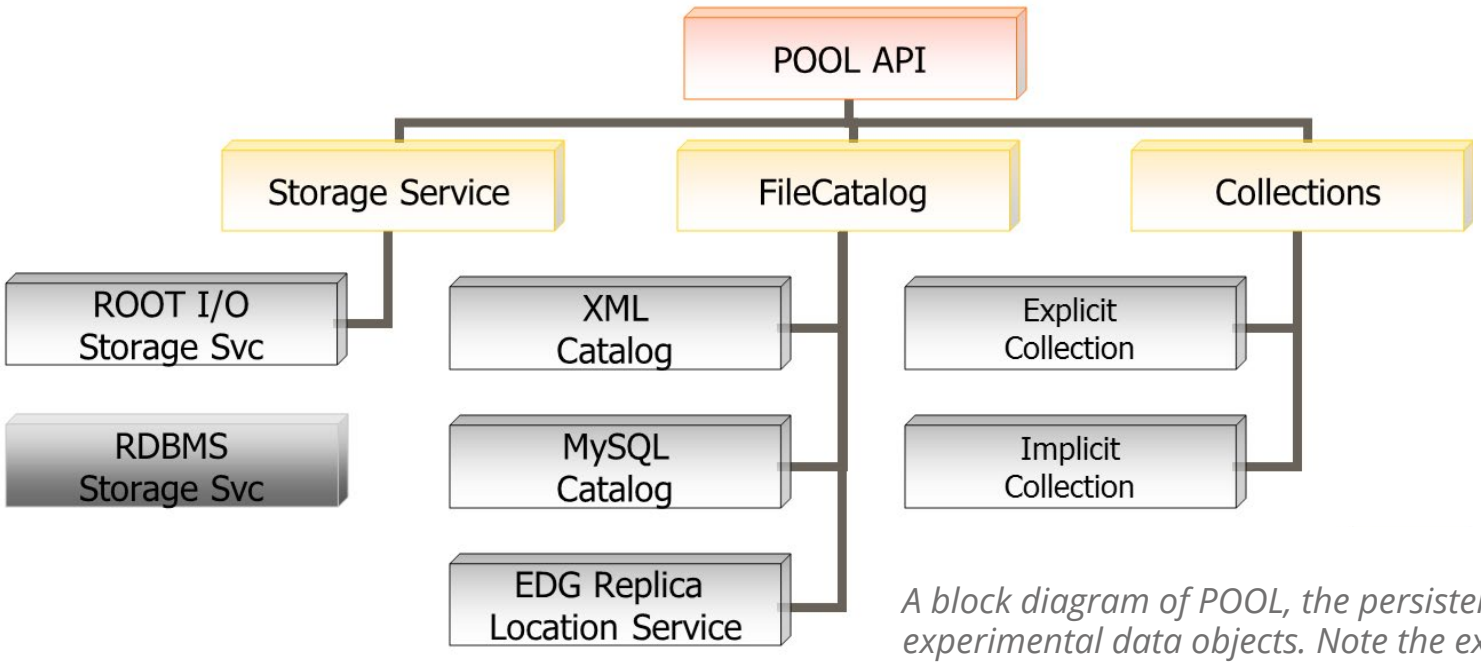


The LHC data are saved on Tier-0 servers. They are then distributed it to 11 Tier-1 centres all over the world.

• Lastly, the NetApp ONTAP 8 system, which enables the addition of storage bays and the displacement of data without interrupting the experiments in progress. This permanent continuity is one of the infrastructure's strong points: it is a condition of maximum availability while the cluster mode makes it possible to fill all kinds of scaling needs.

When CERN started operation, the IT team chose a NAS-type architecture with SATA disks in RAID 1 - a surprising option compared to the more standard SAN FC alternative. However, after several years of intensive use, the statistics speak (by the voice of Eric Grancher): "*Since 2007, we did not have any downtime connected with SATA disks, and we have not lost a single data block. Clearly, the reliability is as good as in FC technology.*"

At the end of this introductory tour, the conclusion is that current technologies enable managing the most complex data and the most enormous volumes. If the case of CERN represents the paroxysm of scientific Big Data, the paths followed are certainly not the only ones possible. But they have the merit of efficiency, a quality acclaimed by all the researchers who have had the privilege of putting them into practice. Some technical issues mentioned along our visit call for a more thorough examination. That is the role of the boxes that complete this story.



A block diagram of POOL, the persistence framework for experimental data objects. Note the extensive use of flat ROOT files, containers around which CERN has standardized its operations.

A unique laboratory for scientific Big Data

Research at CERN has implications that go far beyond basic physics. Remember 1989: it was a CERN researcher, Tim Berners-Lee, who invented the World Wide Web to enable the remote sharing of information between scientists. More recently, the harnessing of particle acceleration gave birth to PET scans found in a growing number of hospitals, and also cutting-edge nuclear material detection equipment.

Today, given the volume of data produced by the various facilities of the Centre, CERN also stands out as a pioneer in Big Data management. This article is not short of numbers, but if there is one that should not be overlooked, it's the 25 Petabytes of usable data generated last year only - and this is after 99% of the information from the LHC detectors was not considered pertinent enough to be stored.

CERN has opened its library of information to researchers from about 150 sites worldwide. Obviously, this amazing feat requires raw experimental data to be stored outside of a relational DBMS. They reside in ROOT files, which are particularly well-suited for scientific analysis thanks to their unique consistency model (write once), while persistence is managed by a custom framework called POOL (Pool of Persistent Ob-

100 Petabytes to maintain

Within the past three years, the volume of data collected from the LHC alone reached 75 PB, bringing the overall volume of data generated by CERN to more than 100 PB. At this level, storage is an experiment in itself. The strategy for CERN's IT directors has been to characterize the information according to its probability of access. Thus, 88 PB have been archived on tape via the CERN Advanced Storage system (CASTOR), while about 13 PB have been stored on a disc system (EOS) optimized for rapid and simultaneous analysis by multiple users.

It is when addressing the logistical aspect that these orders of magnitude take on their full meaning. CERN has deployed eight robotic band libraries, spread over two buildings. Each of them contains approximately 14,000 cartridges with a unit capacity of 1 to 5.5 TB. EOS itself consists of 17,000 discs connected to 800 servers. A unified namespace supports concurrent accesses to the millions of files.

The IT team is taking advantage of the first LHC long shutdown (aka "LS1") to analyze the health of its data treasury and conduct a number of consolidation and maintenance operations. For example, the

jects for LHC). It's the metadata on this treasury that are managed via an Oracle 11gR2 database, therefore in transactional consistency.

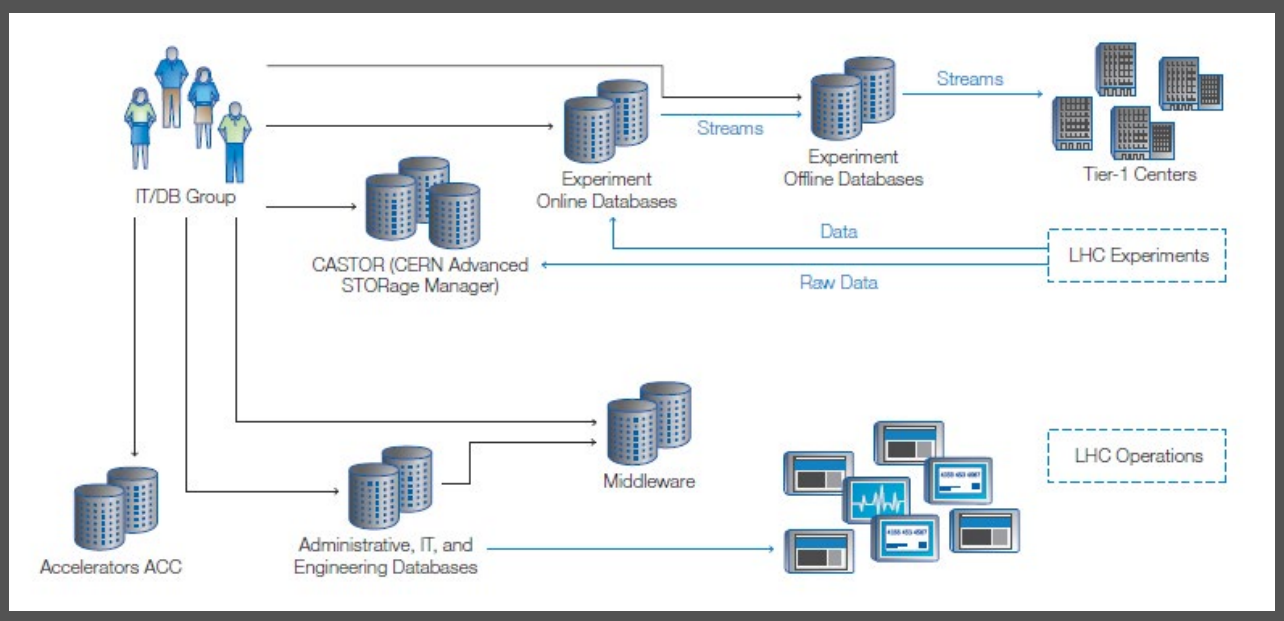
Typically, CERN's raw data is analyzed in batch mode. Everyone knows that this type of processing can quickly become time- and resource-consuming. That is why the Centre's data scientists are working on different ways to optimize not only access but also requests. These efforts are undertaken together with various big players (Intel, HP, Oracle...), within the frame of the OpenLab project (FP7). The specific goal is to accelerate the speed of data partition in such a way that it correlates with the increase in production volumes.

Among the paths studied are NoSQL technologies such as Hadoop and Amazon's Dynamo DBMS, which appear to be the most promising. Why? Firstly because they scale up particularly well, especially in terms of distribution on a *large* number of clusters. Secondly because, as such, they free the team from managing the complexity of relational databases. Thirdly because they lend themselves perfectly to storage on commodity servers – standard equipment being always preferred by CERN's IT managers.

The work going on today focuses on the development of procedures allowing the use of links within the Oracle databases in order to extract pertinent information, to process it within Hadoop or as pure business intelligence (as businesses have been doing for a long time) and then to replace it within other Oracle databases. Returning to Oracle has several benefits among which the robustness of the system, its own functional capabilities and the fact that it is universally known and mastered. Concerning the most commonly cited flaws, namely the required processing and storage capabilities, CERN does not consider them to be insurmountable. Taking into account the decreasing costs of specialized cloud storage, the usability benefit seems justified.

Let's not forget either that the bulk of the LHC data analysis is realized by a global network comprised of more than 150 computing centers (Worldwide LHC Computing Grid - WCLG), which are constantly augmenting capacity. Add to this a budget devoted to the use of external cloud services such as Amazon S3 and the strategy makes sense. We know that one of the major obstacles on the way to the next scale is the speed at which data can be transferred from one server to another. In this respect, researching ways to save time and extend *in-situ* processing functionalities will probably benefit the entire community.

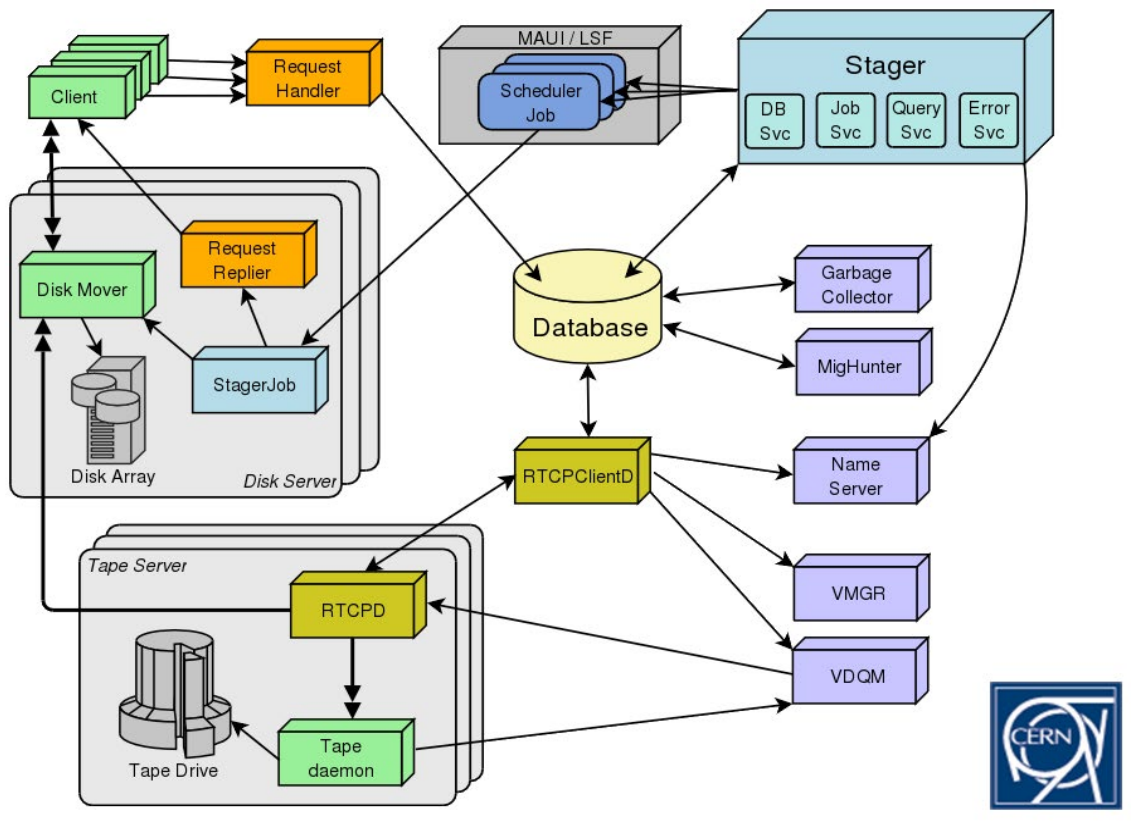
tapes will be replicated on cartridges offering larger individual capacity. At the same time, engineers must prepare for the arrival of new data flows coming from updated accelerators. For this, a new remote datacenter is being created in Budapest (Hungary), with a cluster featuring no less than 20,000 cores and a storage capacity of 5.5 PB. It is currently being interconnected with Geneva via a dedicated 200-Gbps network.



Duality of the LHC information systems: the first one is dedicated to experimentation, the second one to administrative tasks.

A focus on CASTOR, CERN's storage manager

To meet its persistence and access needs, CERN has developed its own hierarchical storage management system - CASTOR (CERN Advanced STORage manager). The initial idea was to enable researchers from around the world to register, list and retrieve files from the command line or application tools using a unique dedicated API. The basic principle of CASTOR is that files are always accessed from the cache drives. It is the back-end of the system that manages their conservation on tape. To help researchers use their existing codes, multiple access protocols are available: RFIO, ROOT, XROOT and GridFTP.



Architecture of the CASTOR system. The role being played by Oracle DBMSes is central. The stager is the port of entry. CERN Document.

At the heart of the CASTOR architecture, a central database backs up the components status changes. *Grosso modo*, this architecture is based on five functional modules:

• a stager allocates and reclaims storage space, controls client access and manages the catalogue of local disc clusters.

• a name server (files and folders) includes files metadata (size, dates, checksums, owners, tape copy information). Management of the namespace is based on command line utilities derived from the Unix model.

• the tape infrastructure provides for the copying of files under certain circumstances (*e.g.* to secure the data) and for the storage of files that are larger than the immediately available disc capacity.

• a client software allows the user to upload and download data files and to manage CASTOR data.

• the Storage Resources Management System gives access to data in a grid environment via its own protocol. It interacts with CASTOR on behalf of a user or of other services such as FTS (the file transfer system used by the LHC community to export data).

To maximize storage efficiency, CASTOR prioritizes large files (of 1 GB minimum). For most small and/or non-experimental containers, users are invited to use AFS, a more conventional file distribution system. Typically, a file stored via CASTOR will be taped in less than 24 hours (8 hours for experimental physics data). Disc instances are then deleted asynchronously.

The duration of reloading data from a tape may vary according to the system's current workload but the average time is about 4 hours. Applications requiring this data will wait from the time of the query until the successful completion of the download, after which they'll start automatically. However, preparation requests can be sent for data to be loaded onto disc in advance. In addition to these procedures, groups of researchers / experiments have an SLA assigned to them that defines precise storage strategies with regard to the work being carried out. ■

INTERNATIONAL
SUPERCOMPUTING CONFERENCE

ISC'14 THE HPC EVENT

June 22 – 26, 2014, Leipzig, Germany

Join the Global Supercomputing Community

www.isc-events.com/isc14

/discover

ETP4HPC:

THE FUTURE OF HPC IN EUROPE IS AT STAKE...

Right now, the decision is being made as to what strategy the "Old continent" will be taking in the race to Exascale and beyond. At a time when the two major blocks of countries either side of Europe are competing against each other in their computing ambitions, the question of European leadership in HPC is more acute than ever. To address it, the program put together by ETP4HTP is aiming to be concrete, overarching and… intelligently funded.

JEAN-FRANCOIS LAVIGNON[1] , JEAN-PHILIPPE NOMINE, PHD[2]

Support the competitiveness of European HPC technologies to support Europe's competitiveness in general - that's how one could sum up the mission of ETP4HPC, the new European Technology Platform for HPC. Established in mid-2012, the organization initially brought together HPC manufacturers and ISVs – Allinea, ARM, Bull, CAPS Entreprise, Eurotech, IBM, Intel, Partec, STMicroelectronics and Xyratex – but also major computing centers and research bodies such as BSC, CEA, CINECA, Fraunhofer, Forschungszentrum Jülich and LRZ. As a platform opened to any organization working in HPC R&D in Europe, or contributing to the development of a European HPC ecosystem, ETP4HPC is also growing itself. As we go to press, the initiative has brought on board another twenty members, and several SMEs are in the process of signing up.

ETP4HPC was born with the publication of its first ***Vision paper***, followed several weeks later by its first official letter to European Commission Vice-President Neelie Kroes. After almost a year of preparatory discussions between the members and with political institutions, ETP4HPC's work has now been set out publicly in its Stra-

(1) Bull - ETP4HPC Chair
(2) CEA - ETP4HPC Office

tegic Research Agenda (SRA): Achieving HPC Leadership in Europe. Producing an SRA is effectively the whole raison d'être for an ETP (see *Understanding European Technology Platforms*). Based on the collective vision of those involved in the sector, it sets out a number of paths to be taken, both in terms of research objectives and its sustainable funding.

Supporting the ecosystem as a whole

While the Vision Paper defines the ETP's objectives and missions by outlining the issues to be considered in the overall R&D program, the SRA sets out a detailed menu for the program by drawing out the main principles for the construction and orchestration of a complete ecosystem. The first difficulty was to define R&D priorities across all HPC technologies. To do this, the ETP members wanted to solicit the opinions of other experts (*see Behind the scenes...*), as well as an expression of user needs from the private sector and from computer simulation software publishers. The latter represent an essential link in the value chain, between technologies and "business" usages, in the vast majority of economic areas. As a result, over 100 contributors were approached in total.

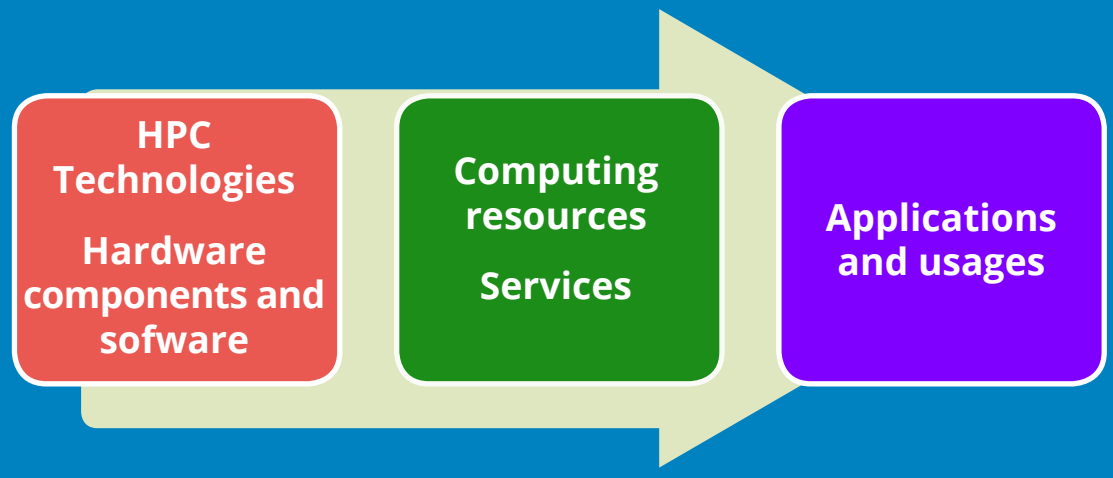
ETP4HPC's interactions with the industry and ISVs explicitly confirmed the importance not just of producing very large scale supercomputers in Europe, but also medium-sized systems that are robust, affordable, efficient and easily program-

HPC in Europe and its value chain

HPC is strategic for research, industry, the economy and our society as a whole. The European Commission formally set out this vision in its February 2012 communication, and following this the Competitiveness Council reaffirmed its willingness to launch a dedicated program on 30 May 2013.

The aim of this program is for the Old Continent to strengthen its position as a major player in the use of HPC, and to use this position to improve its industrial competitiveness and scientific leadership. Europe has all the necessary expertise to design and develop supercomputers and software at the highest level in the world. So it is possible to put in place a comprehensive, sustainable and competitive ecosystem, within the European Union, around HPC technologies.

This ecosystem could be set up by locating its three main cornerstones along a value chain that links technologies to their uses. Based on this vision, ETP4HPC's role is to encourage the development of the necessary HPC technologies in Europe, on a spectrum that spans from hardware components to software tools, and to stimulate the development of complementary services, including access to computing resources, code optimization, etc. In doing so, ETP4HPC embodies the technology providers' response to the expectations and ambitions of the European Commission.



The second cornerstone of the value chain consists of bodies such as the pan-European computing infrastructure PRACE, and institutional and private computing centers who use and implement HPC technologies (providing access to computing resources, associated services...). The third cornerstone – Applications and usages – brings together all kinds of users from the worlds of research and industry, for whom computer simulation and HPC represent increasingly indispensable tools.

In order to effectively fulfill its role, ETP4HPC has already built up the necessary links with other players and initiatives evolving within this ecosystem. This applies, in particular, to the European Exascale Software Initiative (EESI), which is proposing a roadmap to transform applications to operate at the exascale level by the end of the current decade. It also includes a representative majority of industrial HPC users and computer simulation software vendors.

mable. To this effect, the SRA prepared by ETP4HPC is structured according to a multi-dimensional model (**fig. 1**).

A 360° vision

The first of these four dimensions, *HPC Stacks Elements*, represents the evolution and improvement of the fundamental hardware and software elements that make up high-performance computing systems of any size and scale.

The stack of these elements and subsystems is mainly made up of hardware components and architectures (processors, memory, network...), system and management software (OS, configuration management, resources management...) and programing environments (taking into account massive parallelization and with a special focus on programming API standardization).

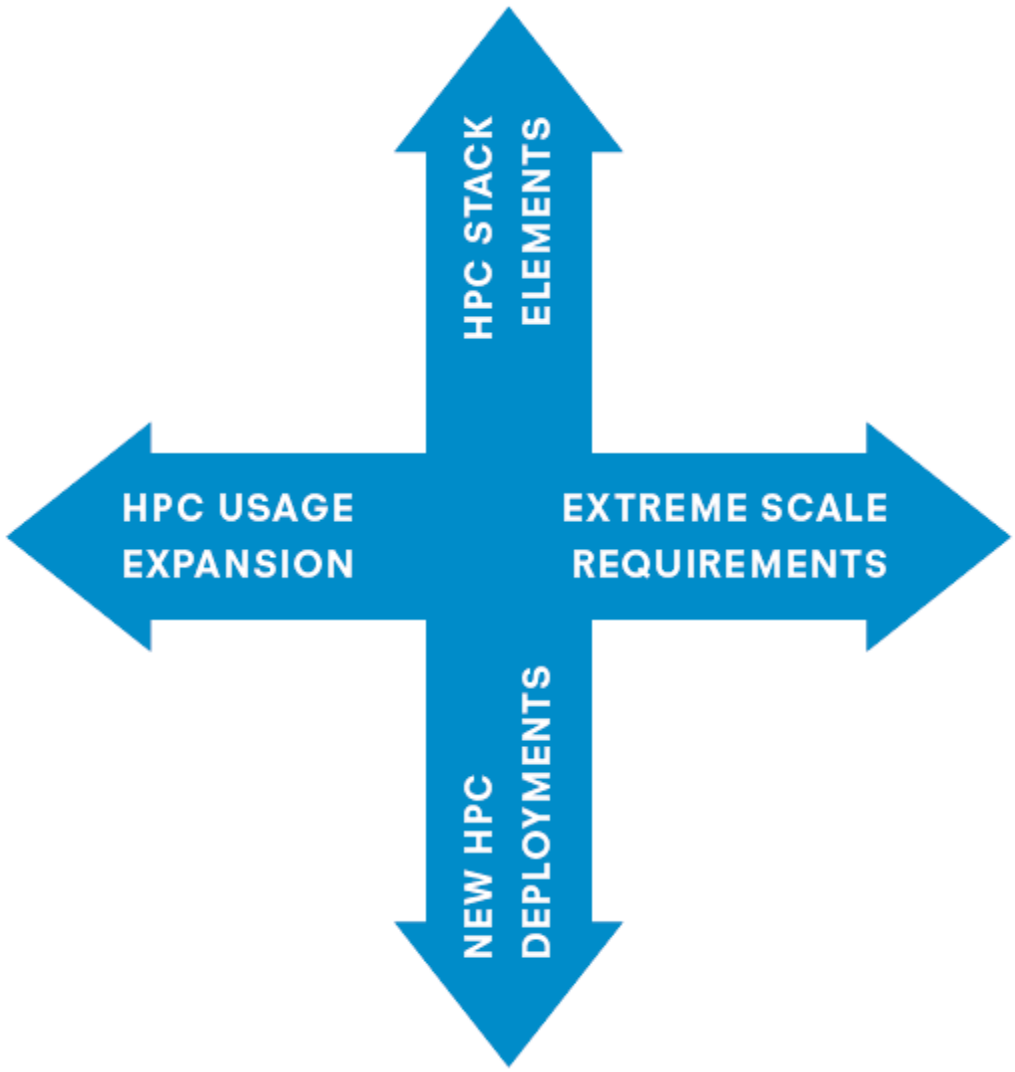


Fig. 1 - Multidimensional approach for the development of HPC in Europe according to the ETP4HPC's Strategic Research Agenda.

The second dimension, Extreme Scale Requirements, identifies the conditions needed to successfully move to Exascale and beyond. The integration of elements from the first dimension at these levels requires a global, transverse vision of the many problems that are exacerbated at such large

Understanding European Technology Platforms

The primary role of a *European Technology Platform* or ETP is to develop and regularly update a roadmap of research priorities in its sector. These *Strategic Research Agendas* feed and influence European research programs and their funding.

In essence, ETPs are driven by industrial companies. Various bodies come together to define an R&D program around a certain number of strategic challenges on a given theme: a theme where the future growth of the Union depends on very large-scale research and technological progress in the medium and long term. This is why ETPs put forward a vision that is shared by all their contributors: industry, public authorities, the scientific community, regulatory bodies, civil society, operators, users...

Launched to coincide with the Sixth European R&D Framework Programme (2002-2006), ETPs have taken on growing importance under the Seventh Framework Programme, which covers the period 2007-2013. Although they remain independent structures whose mission is to make recommendations that are recognized by the European Commission as direct interlocutors. In certain cases, ETPs lead to the creation of operational structures that directly implement research programs, such as Joint Technology Initiatives (JTIs) or Public Private Partnerships (PPPs).

There are currently about 40 ETPs today, in many different domains: information and communication technologies (ICT), energy, bio-economics, ERP, transport... In ICT, they include ENIAC (nano-electronics), ARTEMIS (embedded computing systems), EPOSS (intelligent systems), NESSI (software and Internet services). ARTEMIS and ENIAC are among the few ETPs that have led to the creation of a JTI.

A year of intense exchanges between experts, industrial users and major players in the industry was required to establish a comprehensive program to strengthen the position of Europe on the HPC international scene.

scales: energy efficiency, resilience, balancing performance between computation, communications, storage, etc.

The third dimension, new HPC Deployments, includes new HPC usages – the Big Data, the Cloud, embedded computing, real-time... – and defines ways of stimulating and supporting them. This dimension is distinguished by its highly technical content. In effect, various ambitious developments are needed, in particular when it comes to algorithms (data processing, extracting meaning and value...) and system tools (managing remote access modes, resources virtualization, workflow management...).

The fourth dimension, *HPC Usage Expansion*, is the one dedicated to the development and democratization of HPC usage. Here we leave the purely technical realm to look at equally important actions, to do with the economy and education/training, amongst other things. These actions require co-ordinated relationships with other players, with a view to developing the overall ecosystem mentioned earlier. First and foremost, it will be about facilitating access to various levels of computing resources, promoting their use and optimizing the costs of purchasing and ownership. In parallel, the SRA recommends stimulating the development of a services sector (porting, optimization) which would be particularly useful for ISVs and industrial users, as well as supporting SMEs working on HPC technologies. Finally, it seems essential for all those involved to contribute to HPC training and education. This is area,

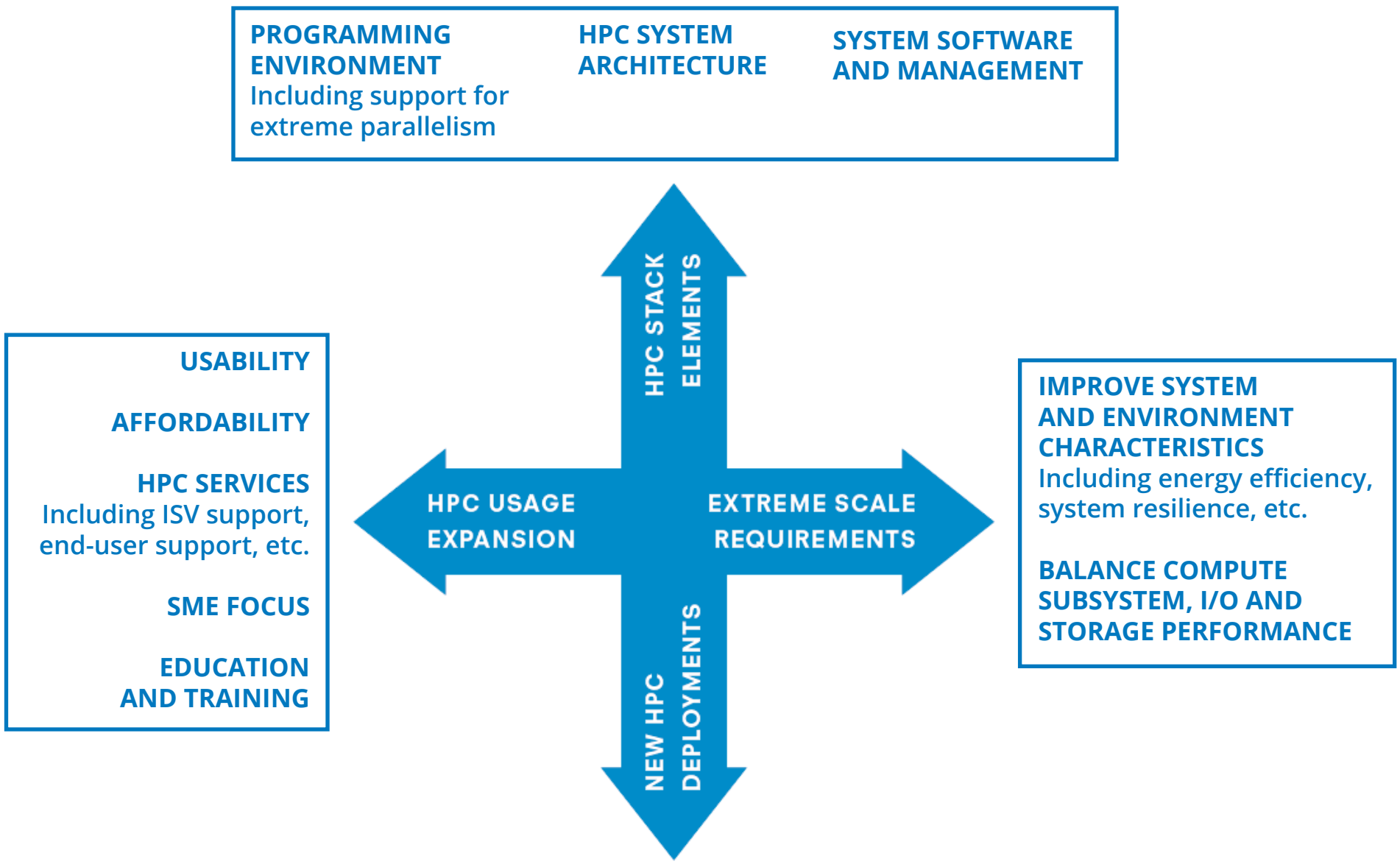


Fig. 2 - From the multidimensional vision mentioned above, the ETP4HPC SRA identifies the major technical areas leading to coordinated action plans. Ultimate goal: that the whole be greater than the sum of its parts.

training programs focused on HPC usage are seen as equally important as those focused on HPC technology.

A detailed plan for shared funding

In the production of the SRA, it seemed relevant to break down the technical aspects of the first three dimensions into six distinct sections, as illustrated in **fig. 2:**

- Components and system architectures
- System / management software
- Programming environments
- Energy and resilience
- Balancing performance between computation, communications, storage
- Big Data and HPC usage models

Based on this categorization, each section is subject to a more acurate analysis and is further broken down into subthemes and research priorities with different milestones (see **fig. 3** next page). A total of 140 milestones are proposed, as well as an overall plan linking them together, with two major phases: Acquisition of the relevant technology capabilities in the various areas, with associated prototype demonstrators (2014-2017), and Consolidation, extension, exascale integration (2018-2020).

This program would be of little use, however, without a realistic estimation of its costs. Excluding the prototypes, this

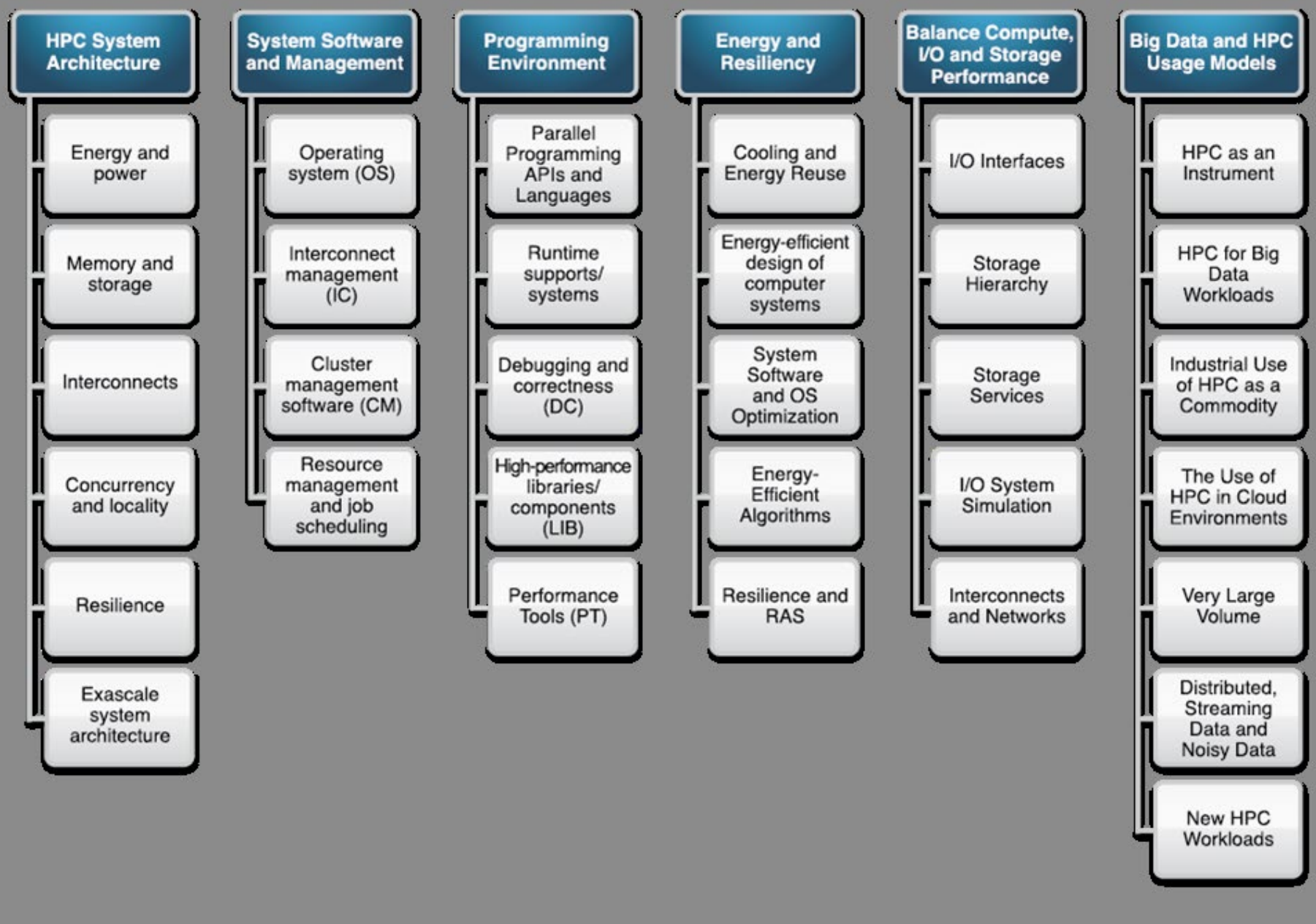


Fig. 3 - Breakdown of technical sections into sub-themes and research priorities.

will be around €150 million a year for seven years. The European Commission will be approached to fund at least half of this, if possible with a subsidy of over 50%, most notably for the SMEs involved. The rest of the funding will come from industry and research organizarions who will underwrite associated projects and will be retained to carry them out.

At the time of going to press, the implementation of the program has not been fully confirmed nor finalized. ETP4HPC's message is that if the European Commission offers support at the proposed level, the industry will be able to commit to the proposed R&D program, with performance indicators and monitoring metrics to be defined along with the mechanisms for implementing the whole structure.

In the face of likely developments in the USA and the growing power of China in HPC, this is no small challenge. Europe has the resources in keeping with its ambitions. Now it has to work out how to mobilize them. ■

About the platform:
http://www.etp4hpc.eu

Behind the scenes...

Throughout 2012, meetings held every month by the Steering Board (drawn from the 15 founder members) led to the creation the ETP4HPC association, the definition of its statutes and the setting up of an Executive Office. The Office runs using personnel shared across four satellites in France (CEA), Germany (Partec), Italy (CINECA and Eurotech) and Spain (BSC). It is responsible for supporting the association's overall operation, its promotion and communication, as well as the development of the SRA. The main steps that it have been carried out are summarized below:

March 2012:
- ✓ Publication of the *Vision paper*

June 2012:
- ✓ Letter to European Commissioner Neelie Kroes announcing the creation of the ETP
- ✓ Public presentation of ETP4HPC at ISC'12 (Hamburg)
- ✓ Work starts on the SRA; first working seminar of experts in Paris

September 2012:
- ✓ Steering Board meets with Neelie Kroes in Brussels: positioning and discussion of the economic impact of HPC, ETP4HPC's goals and the advantages of a European HPC policy

October2012:
- ✓ Working seminar on the SRA in Barcelona, in two sessions: meeting with a panel of 15 industrial HPC users, followed by a working session for experts
- ✓ Third working session for experts at the end of October in Bologna

November 2012:
- ✓ Intermediate document between the Vision paper and the future SRA – *Europe Achieving HPC Leadership* – submitted to Neelie Kroes
- ✓ ETP4HPC panel at SC'12

December 2012:
- ✓ Letter of recognition and support for the ETP received from Neelie Kroes
- ✓ Fourth and final working seminar for experts in Munich, working towards the SRA
- ✓ Final incorporation of ETP4HPC as a legal entity (under Dutch law)
- ✓ Interview with a panel of computer simulation software vendors, working towards the SRA

February 2013:
- ✓ SRA finalized

April 2013:
- ✓ Publication of the final SRA
- ✓ Preparation of the officially recognized list of ETPs by the European Commission
- ✓ ETP4HPC formally declares its existence

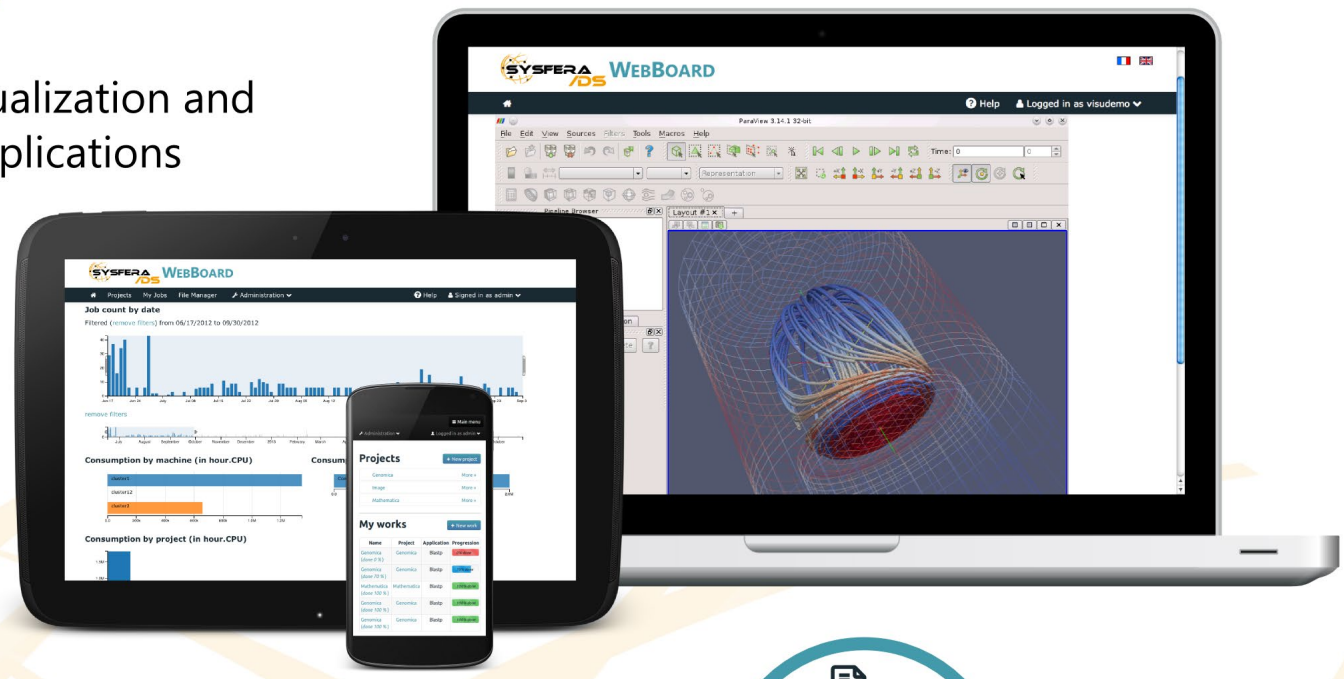
May 2013:
- ✓ EU Competitiveness Council of May 30th recognizes the importance of HPC as a whole

June 2013:
- ✓ ETP4HPC submits an expression of interest for an HPC PPP (Public Private Partnership) to the Commission
- ✓ ETP4HPC stand and event at ISC'13

Manage your hybrid HPC environments and visualize applications with a web interface

- ✓ User-friendly, collaborative scientific portal
- ✓ Unified and mutualized computing resources

- ✓ Remote visualization and usage of applications



- ✓ Statistics and reporting for billing resource usage

- ✓ Management of distributed data and files

Join us on booth #2417

SC13 Denver, CO 2013

SYSFERA HPC Made Easy

13, avenue Albert Einstein – 69100 Villeurbanne, FRANCE
Tel. : +33.4.81.76.28.90 – Fax : +33.9.85.86.92.55

www.sysfera.com

Verbatim

ALBAN SCHMUTZ

VP, BUSINESS DEVELOPMENT, OVH

AN INTERVIEW BY ALEX ROUSSEL

On the agenda:

- The acquisition of cloud HPC specialist Oxalya
- A 5 Tbps network and a million servers capacity
- OVH *vs* AWS
- Offering Simulation as-a-Service

Mr Schmutz, for the people of the Big Data and HPC community who would not know you yet, what is your profile and what are your responsibilities today at OVH?

Alban Schmutz : I'm in charge of OVH's business development relations with our partners, with vendors and developers, as well as with institutional organizations and the general public. The most recent illustration of this is the appointment of OVH as copilot with Atos on a cloud computing project for one of the 34 industrial plans announced by the French government.

> *FOR A RESEARCHER, WHETHER PUBLIC OR PRIVATE, IT IS MUCH MORE INTERESTING TO WORK ON THEIR SPECIALTY THAN TO MANAGE COMPUTER PROBLEMS...*

Historically, I started my career creating an open source software company - Linagora. This name is now well-known and has some authority in the area. I then left Linagora to create Oxalya, which began operations in 2005. Oxalya specializes in HPC on demand with a strong focus on numerical simulation. This is why it was bought by OVH in 2012.

Exactly what made you create Oxalya and how did this lead you to the cloud?

Several customers came to Linagora asking us to develop their HPC infrastructure. The reason was obvious: a computation cluster, it's Linux in 95% of cases. And when you are a specialist in open source software, people come to you for this type of sensitive matters. At the time, quite frankly, we didn't know much about supercomputing. But we did it for a first customer, actually a university.

When we got a second request of the same kind, we thought we could probably automate the process. That's how Oxalya was born: rather than having to do the same thing for each client all over again, why not propose solutions based on a real added value of automation? You know, typically, we're dealing with scientists doing system administration, whereas their talent is in research. It is much more interesting for them to work on their specialty in physics or chemistry than to manage computer problems. Oxalya's vocation is to make life easier for supercomputing users by automating the management of their infrastructure.

Were you already thinking of cloud computing in 2005 or were you still concentrating on on-premise projects?

Initially, Oxalya only worked on on-premise projects. We were well aware that people wanted to have their machines locally but, at the same time, we thought that pooling resources made more sense. The use of resources will always fluctuate. Load peaks exist, but overall usage changes over time so that the average usage rates are never optimal. It is therefore more appropriate to share. However, the use of computing clusters outside of the organization, public or private, poses a number of problems: how do I remotely access my resources, view data, work in collaboration with other researchers based on other sites and so on? This distributed approach, this "HPC cloud" that we initially had in mind - the market did not seem ripe enough to adopt it. But finally everything went very fast. From our first on-premise deployments in 2005, we began to work on issues such as remote visualization and collaboration in 2006, which led us to launch our first "HPC on demand" offer in February 2008...

How did you manage to develop a sustainable business model in a market where initial investment is so high?

We worked with HP to implement our first computing infrastructure. HP helped us with the financing of the equipment. From our side, we had largely financed the development of the solution components. Technically, we had automated management so we could change

the machines in the cluster as we went along and isolate groups of machines for particular clients. Public Tier-1 or Tier-2 HPC datacenters have no problem isolating users, especially in terms of security. On a shared hosting, you cannot imagine having GM and Ford on the same machine. So we worked hard on these problems in our software stack and, as early as 2008, we were ready to go.

At the time, cloud computing was not what it is today, especially in supercomputing. How did you launch your offer?

With a campaign of free cluster hours for SMEs, under certain conditions and for a limited number of users. SMEs typically have budget constraints and skills problems, so the targeting was relevant. This operation was not comparable to free hosting: we really aimed at helping companies to do intensive calculations and we wanted to help promoting their innovation projects. We therefore set quotas and gave millions of hours to the projects that met the criteria.

Was it successful?

Actually... yes and no. Yes in the sense that all our resources were quickly exploited in full. By paying clients - generally large accounts familiar with HPC who only needed machine time - and by the free program clients. So yes, overall, it was successful. Meanwhile, we learned a lot about HPC access for SMEs. A year later, we compared the applications we had received and their effective use... and we saw there was a big difference. Among the companies that did not use the platform, in approximately 25% of the cases, the person making the request was gone. Another quarter of the companies who had asked for it did so because it was free, but had no real computing need.

It was all the more surprising as serious applications files had to be completed and approved by a technical committee. We also noticed that some of the requests were made by students... But the really interesting point is that the SMEs who truly used our resources were already convinced of the strategic aspect of supercomputing. Clearly, they wanted to be accompanied by an external service provider to save time and be more efficient.

> "OVH IS NOW THE MOST IMPORTANT INTEL CUSTOMER IN EUROPE. WITH A PARTNER LIKE THIS, OXALYA HAS THE CRITICAL MASS TO PROVIDE MORE THAN AMAZON...

It is a little over a year that Oxalya was acquired by OVH. What brought you to sell to a web hosting company?

Oxalya and OVH are two complementary strategic businesses. Oxalya's problem was investment. OVH's problem was to find a path to grow in sectors that were not necessarily theirs. Today, OVH is the most important Intel customer in Europe. When OVH's potential HPC growth came into the discussion. Intel told them that if they wanted to do HPC, they should tell Oxalya. In the HPC community, Oxalya has always had good press.

From our side at Oxalya, we were knowing the guys from Intel quite well, since we were selling a lot of Intel configurations to our HPC customers. Intel also knows us well from our innovative components. At the time, in remote visualization for numerical simulation, Oxalya was the only one ahead. Take for instance the image wall deployed by EDF R&D. The wall part was supplied by Barco but the software and infrastructure part was provided and implemented by Oxalya. It is our software that allows the operation of 16 synchronized projectors on a slab of glass that weighs two tons. We also managed and coordinated large collaborative research projects, of up to 18 partners. Anyway, to make along story short, Intel knew us well. For Intel, HPC is a growing segment and a prestigious area to demonstrate technology. So they naturally favored the merger.

You didn't think you could continue to grow independently?

Our business logic was software, not hardware, development. We wanted to develop SaaS-enabling tools, to create the software infrastructure capable of hosting other software in SaaS mode. We know how to "SaaSify" an application by installing it locally and making it easily available. Whatever the product from Ansys, ESI, Dassault Systèmes or other vendors, we can do it. We truly master the capacity to automate all the availability processes.

But Oxalya's vocation wasn't to invest in a physical infrastructure. We did it once because we needed to, knowing perfectly well that hosting is a mass market. In my opinion, HPC is a bit limited in terms of critical mass to absorb the necessary investments. So, in partnering with a player like OVH, Oxalya got the critical mass. And it was very interesting for Oxalya to lean on OVH considering the commercial opportunities and scalability issues that could follow. OVH manufactures its own servers, on its own assembly lines. If the need arises, we can have, in a few days, 300 additional servers designed to our specifications and plugged into the datacenter. No need to wait 6 to 8 weeks before a manufacturer supplies us.

OVH is experiencing a disruption in dedicated servers, following a successful commercial offer. Is this not a limitation of the model and, more importantly, what was the impact on the HPC activity?

Again, yes and no. There is currently a sold out on part of our offers but we are not really out of stock on servers. In fact, we want to change our business model, and that is why we stopped the orders. We need to take some time out to ask ourselves the right questions, to put in place the proper mechanics. Then we'll reopen the valves. Now if a HPC customer calls us for a large amount of servers, of course, we can deliv-

er. Our supply chains averages between 500 and 1000 servers a day. With a capacity this large, things can go very fast. On an order of several thousand servers, we won't be ready in 24 hours but then again, in HPC, you never have this type of 24 hours delivery request. However, in a few days, it is quite possible. And in software terms, we have the infrastructure to scale at will.

What capacity do you still have in your datacenters?

We now have 170,000 servers in production, but we have the infrastructure to produce up to a million. With this almost unlimited capacity, we can move very quickly. For Oxalya, this is a very important differentiating asset. A pure player of HPC will now have a hard time catching up, because building a datacenter is complicated but also because energy is the primary cost of a datacenter. OVH has a PUE of 1.09 in its two centers at Gravelines in France and Beauharnois in Canada. And yet, this figure is calculated in the worst case way, *i.e.* directly out of high-voltage lines. We prefer to use this figure since PUE is not a standard yet. By changing some of the measurements we could still go lower.

Moreover, OVH is now present in 16 countries. In this respect, our new Canadian center is an ideal gateway to the North American market. Plus, we also have our own network - based on a huge 5 Tbps backbone! If one of our HPC customers experiences network congestion, it will be from their side, not ours. And once the data is with us, remote visualization takes over. These networking capabilities increase the value of everything that has been developed so far on the software side.

What is the advantage of having your own interconnection network between your datacenters and access points ?

In terms of SLAs, it is very important to have a network. We can measure our quality of service between our gateway in Hong Kong and our center in Canada. And what we can measure, we can commit to. When you use a standard cloud, you go through an operator network to access it. Then you go in multi-operators mode up to the destination. It's a best effort route for everyone, which makes offering an end-to-end SLA impossible. I personally worked on the ETICS project whose goal was to implement end-to-end SLAs among a group of between 15 operators. I won't go into details here, but it just can't be done! For operators, the availability of their networks and their various routes is core business. They are certainly not going to offer competitors the ability to manage their own routing. It is therefore impossible to commit to an end-to-end SLA today except if you have your own network, which OVH does. This is a huge differentiator.

> ***PRESENT IN 16 COUNTRIES, OVH OWNS ITS OWN GLOBAL INTERCONNECTION NETWORK, WITH A 5 TBIT/S BAND-WIDTH AND AN UNPRECEDENTED QUALITY OF SERVICE...***

OVH is the number one hoster in Europe, and number three worldwide, but is it big enough to compete with Amazon Web Services, the undisputed leader in cloud computing, especially in HPC instances?

We can definitely compete on quality of service and pricing. OVH is significantly cheaper than Amazon in almost all ranges, not only in HPC. On a comparable offer, OVH is 30% more attractive. If we can beat the rates, it is because we have the critical mass, a great PUE, our own very high bandwidth network and an integrated design chain. All these aspects allow us to offer excellent service quality. We can also compete on the security part, on data location. For example, for our French customers, we can guarantee that resources and customer data will remain located in France, while Amazon has no datacenter in France. Given some recent events, it is a problem.

Regarding HPC infrastructure more specifically, we can resize at will because our cloud instances are not designed like Amazon's. OVH gives root access to customers, so they can manage their servers as they see fit. In addition, we have on-demand jobs that include almost all numerical simulation tools. That will not be available on Amazon anytime soon.

All in all, there are many points on which we are very different from Amazon. Oxalya is a pure player: our solutions were designed from the start for numerical simulation users. With Amazon, it's rather you install your virtual machines locally and then... off you go! Of course, we also offer self-service machines, with NVIDIA-equipped accelerators and so on, but we can provide very finely tuned systems for intensive calculations or configurations specifically designed for particular uses. Amazon does not have this.

In terms of security, OVH also innovates quite uniquely. Take DDOS (Distributed Denial of Service) attacks for instance. With our infrastructure, we can take 500 Gb/s of traffic and keep the capability of sorting entry points. We know how to establish multipoint VPNs, we can provide levels of security specific to a given environment...

When Oxalya was on its own, we already used these technologies in HPC, for example to isolate InfiniBand flows. We worked hard on this InfiniBand problem, on the virtualization and network deployment automation. It is this a very specialized know-how that OVH offers globally today.

Wasn't there a marketing risk for a specialist such as Oxalya to be absorbed by a known heavyweight in the web hosting business?

(Smiles) The question we faced was rather "Who is OVH?". Our customers are people who buy HPC infrastructures - generally not CIOs or web players. Besides, even among CIOs, not everyone knows OVH necessarily. When we told them what I just mentioned, the message was clear: we've jumped from a few hundreds to hundreds of thousands of machines. This real change in scale does not mean that Oxalya is drowned in the OVH structure. We kept all our contacts in the HPC community and therefore, on the contrary, it only strengthens and broadens our offerings. There is now a larger structure behind Oxalya but the brand continues.

For how long?

Difficult to say. At OVH, everything evolves very fast. Last year, it was decided to keep it, and a year later it is still there. We will see what happens, but the hosting and HPC markets are well differentiated, so keeping the Oxalya brand makes sense.

How did the Oxalya activity develop since the takeover? Do you have any figures in terms of customers, of HPC servers at OVH, for example?

We don't disclose any business figures. We never communicated on this point and we do not want to do it today, all the more as we are still in launch phase. Over the past year, we focused our efforts on adapting the Oxalya software stack to the OVH platform. We also standardized Oxalya's offers to the OVH infrastructure. That is basically the summary of our technical activity in the last 12 months. We just launched on-demand HPC "by OVH", with OVH billing, so the current numbers would not be very significant.

But do you at least hold a significant share of the cloud HPC market in the countries where you operate ?

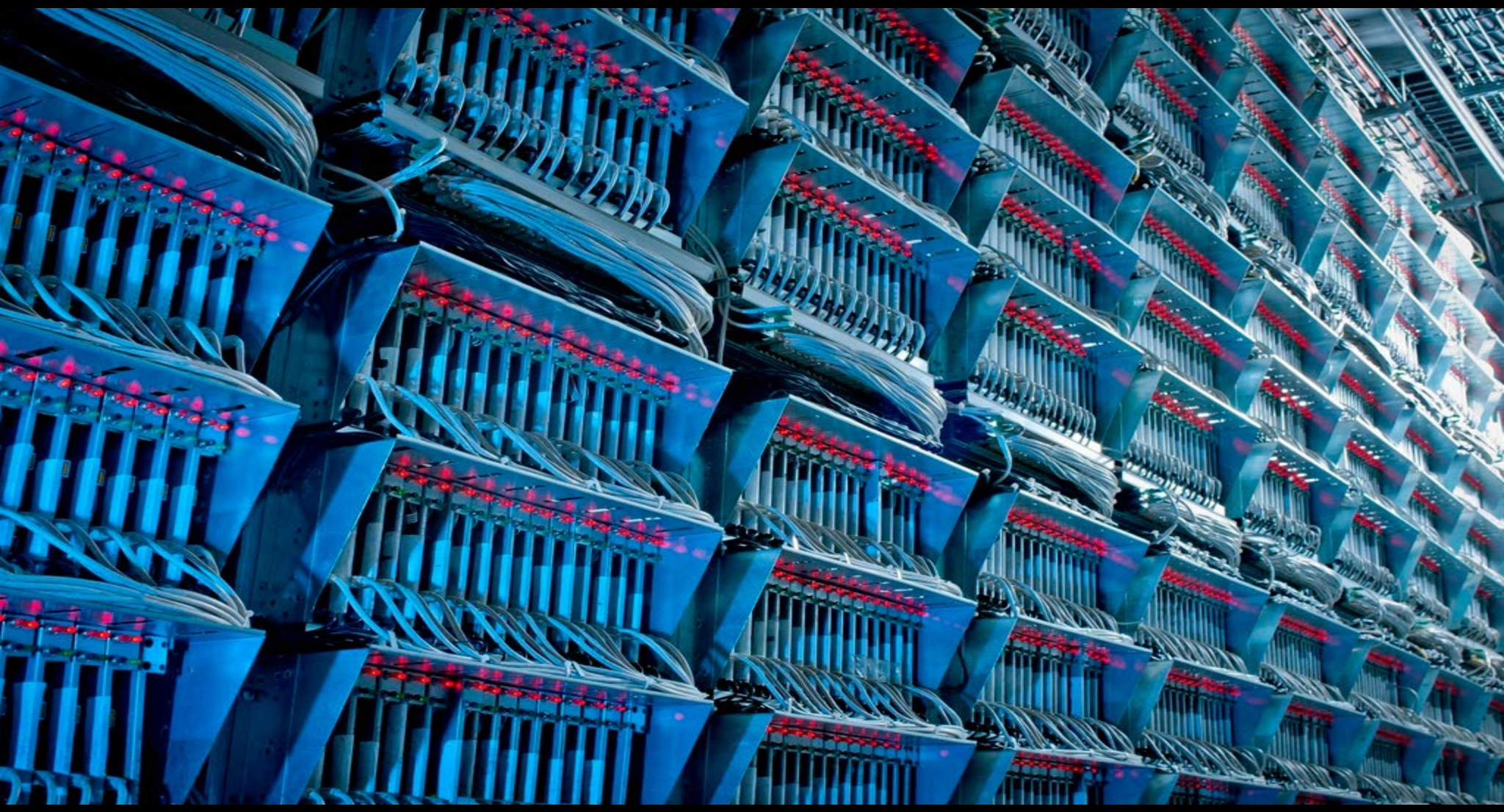
On-demand HPC is not a big market yet. It is still in its infancy, so speaking of market shares does not really make sense. With hindsight, I can say that when we started business in 2008, it was too soon. The situation is a bit strange, but there may be an explanation. The primary audience of HPC are people who work in innovation, academic research, industrial R&D. Paradoxically, this is a very conservative environment, at least from a technical standpoint. When I started Oxalya, my idea was that we could do HPC just like we do Web: see and work on the data online, receive a text or an email when a result is ready, etc. I was wrong. HPC users are very concerned about the location of their data, and everyone considers their data to be more strategic than others'...

Is simulation data more strategic than CRM data? For a certain number of companies, I'm not sure. The ordinary R&D culture is that everything must be partitioned and done at home. Opening is difficult, and this is where the paradox lies. At the same time, as people consume more and more online services on their smartphones and tablets, they end up thinking that eventually they could also use HPC this way. What we do today, we were able to do it five years ago. This is an environment that evolves very slowly and with great effort.

But I believe the start is taking place now. The market is ready, if only because the shifting of costs from capex to opex is interesting almost all businesses. Everyone needs to streamline and, in this context, we provide additional capabilities that are both flexible and low cost. Also, global offers have matured. Suppliers have now reached a certain size, like Oxalya with OVH. Five years ago, they were only small players. What we see today, really, is that the market is going to scale up.

Has your clients profile changed over time? Do you see more SMEs coming to HPC?

No, I do not see any profile changes for now. Today, our natural customer is a large account. But things can evolve. Originally, Oxalya was a B2B company with a very traditional approach to business: go see people, talk to them, study projects together, and then maybe get a chance to sign a contract. It was our culture. For its part, OVH is very industrialized: 90% of OVH customers or-



der through our websites. This is something that will be new for Oxalya customers.

Where are you in your discussions with software vendors?

ISVs will be a growth driver for Oxalya and the entire OVH group. Today, we are working to implement infrastructures for big services integrators who work on complex projects. Our role is not to position ourselves on the business needs: it is the integrators, to whom we provide infrastructure, who lead the projects. So naturally we are doing the same with the ISVs that come to SaaS.

Does OVH offer the technology developed by Oxalya to "SaaS-ify" their applications?

Yes, even though we are positioned in different HPC areas. Our technology is not specific to simulation, we can use it with any application. Therefore, we can work with any software vendor who want their products available in Web mode in addition or in complement to classic usage. This is where we bring added value, by providing a secure, reliable and high performance infrastructure.

What simulation applications are you going to make available?

For HPC in SaaS mode, we will have stand-alone application accesses *à la* remote desktop kind. Of course, in our new series, we offer mainstream open source HPC solutions like Code_Aster or OpenFOAM. But at the same time, we are working with a number of ISVs to allow their users to connect with us using their software tokens.

So you solved the difficult issue of usage-based pricing?

You know, to tell you the truth, some of our discussions have been going on since 2008. At the time, many ISVs would find the project "great". But when we arrived with potential customers, most eventually found that it was "complicated" - the on-demand licenses and all that... So we decided to take another approach where we would be able to use our clients' tokens with on-demand clusters for the requested job. Technically, there is no problem. Then ISVs have to play along. It is still too early to know what the next milestone will be. We have some quite privileged relationships with some of them, even joint projects, but as I said earlier, it takes time for things to happen. Let's talk about it again in a few months... ■

Step up. Scale out.

Introducing IBM NeXtScale System.

The rising demand for intelligence from increasing data volumes, and need for greater efficiency in the cloud, may leave today's data centers inadequate for your requirements. Introducing IBM NeXtScale System™– an easy-to-deploy, cost-effective, hyperscale computing platform that focuses on maximizing density, performance and efficiency for lower operating costs. Its simple and open design integrates with your existing infrastructure and has the capability to help reduce onboarding time by 75% with optional IBM Intelligent Cluster.[1]

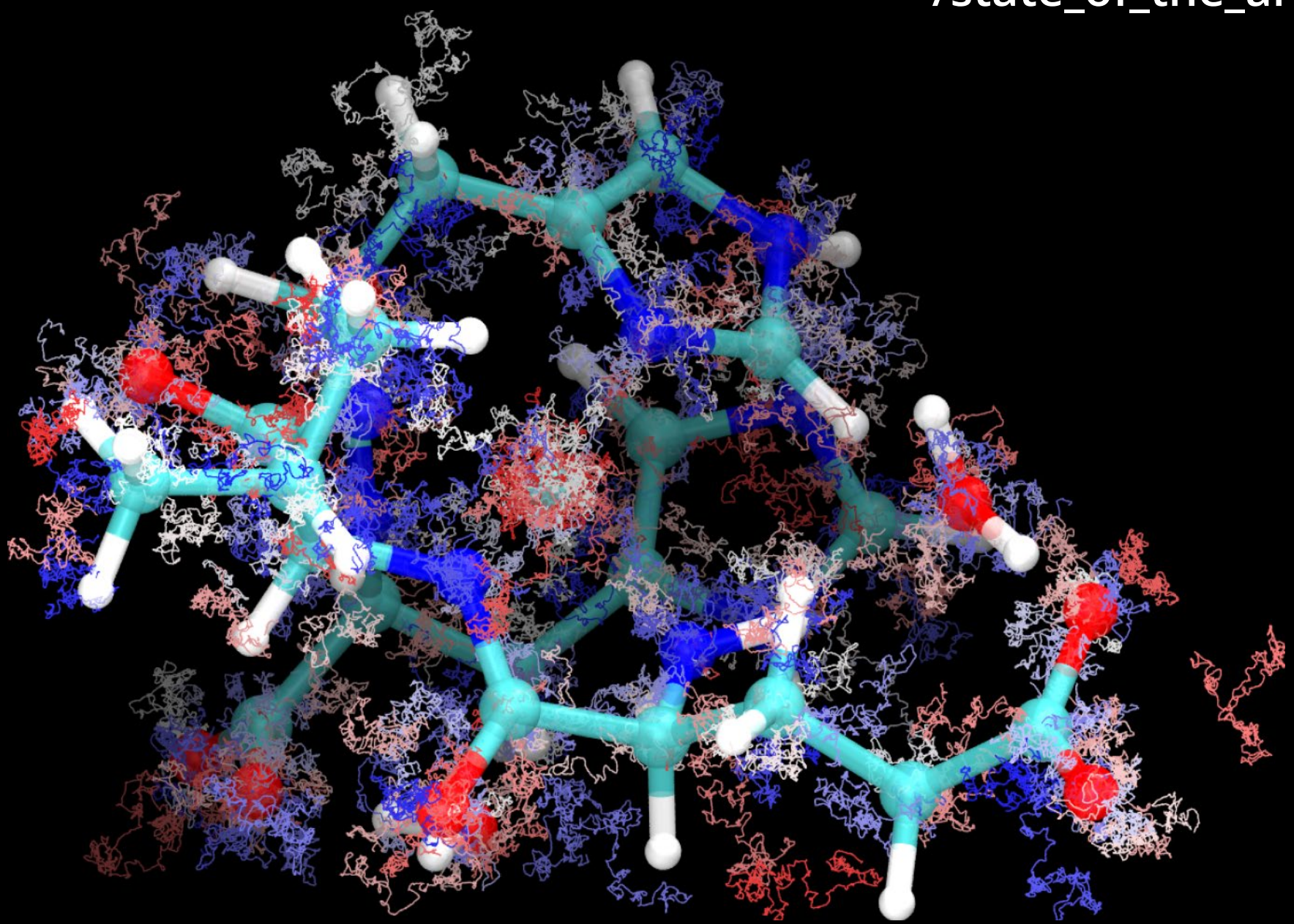
Powered by the new Intel® Xeon® processor E5-2600 v2, IBM NeXtScale System packs 3x the cores[2] versus previous generation 1U rack servers, and up to 37% greater performance[3] and 36% better energy efficiency[4] versus previous generation systems.This high performance system allows you to obtain maximum value from your data by bringing IBM's high performance computing experience to work for you.

See how IBM NeXtScale System can help you optimize your data center for compute-intensive workloads.

Download the Clabby report at ibm.com/systems/nextscale

[1] Based on IBM comparison of configuration and setup by customer onsite versus delivered by IBM using Intelligent Cluster service which is an optional feature available at additional cost.
[2] 3x cores based on industry standard 42U rack comparing 42 1U x3550 M4 rack servers with 2 Intel® Xeon® E5-2600 processors 8 cores each = 672 cores vs 84 NeXtScale System nx360 m4 nodes with 2 Intel Xeon E5-2600 v2 processors 12 cores each for 2016 total cores.
[3] SPECint_rate_base2006 - 669 on IBM iDataPlex dx360 M4 (Intel Xeon E5-2690) vs. 918 on IBM NeXtScale nx360 M4 (Intel Xeon E5-2697 v2.) www.spec.org. Results current as of 9/20/13.
[4] SPECpower_ssj2008 - 5392 on IBM iDataPlex dx360 M4 vs. 7347 on IBM NeXtScale nx360 M4.www.spec.org. Results current as of 9/20/13.

/state_of_the_art



SIMULATIONS IN CHEMISTRY:

THE BENEFITS OF QUANTUM MONTE CARLO METHODS

The question is no longer a debate: scaling up demands new algorithmic strategies. In chemistry, an original approach, the quantum Monte Carlo method or QMC, massively exploits the intrinsic parallelism of probabilistic methods. A savvy implementation can even add real resilience to material failures...

MICHEL CAFFAREL, ANTHONY SCEMAMA*

The availability of massively parallel computers gives rise to new paradigms and opens new routes to numerical simulation. Rather than working hard to parallelize as many algorithms as possible, which do not necessarily lend themselves to the exercise, it can be beneficial to give up the usual approaches and switch to alternative methods. Methods intrinsically inefficient on computers with a few thousand processors but whose algorithmic structure will take advantage of a potentially very large number of processors. We all know that on a single processor architecture, the optimal algorithm is the one that makes it possible to calculate a given quantity in the smallest possible number of elementary operations (for the sake of simplicity, we'll ignore possible constraints related to

* CNRS - Laboratoire de Chimie et Physique Quantiques, U. Paul Sabatier, Toulouse, France.

I/O and memory limitations). The restitution time - *i.e.* the time the user waits to obtain the desired result - and the execution time are then essentially identical and proportional to the number of operations to be performed.

In a parallel architecture, the concept of number of operations to be performed becomes secondary, the objective being to reduce the restitution time through parallel calculations, even if this means performing a much larger number of elementary operations. When the restitution time can be made inversely proportional to the number of computation cores used, we are then in a situation of optimal parallelism. When this situation can be extended to arbitrarily large numbers of processors, we can then rightfully speak of methods that scale up.

As supercomputers are bound to come with an ever greater number of processors, having an algorithm that scales up becomes both a real challenge and a reasonable goal. It is the guarantee that there will always be a minimum number of processors for which the algorithm proves superior to any other that does not scale up, whatever its intrinsic performance (single processor).

Virtual chemistry

The problem is well exemplified in chemistry, a scientific domain extremely demanding in numerical simulations. Advances in drug design, new materials (nanosciences) and renewable energies put it at the heart of our daily lives. Being able to understand, predict and innovate in this field is a considerable society issue.

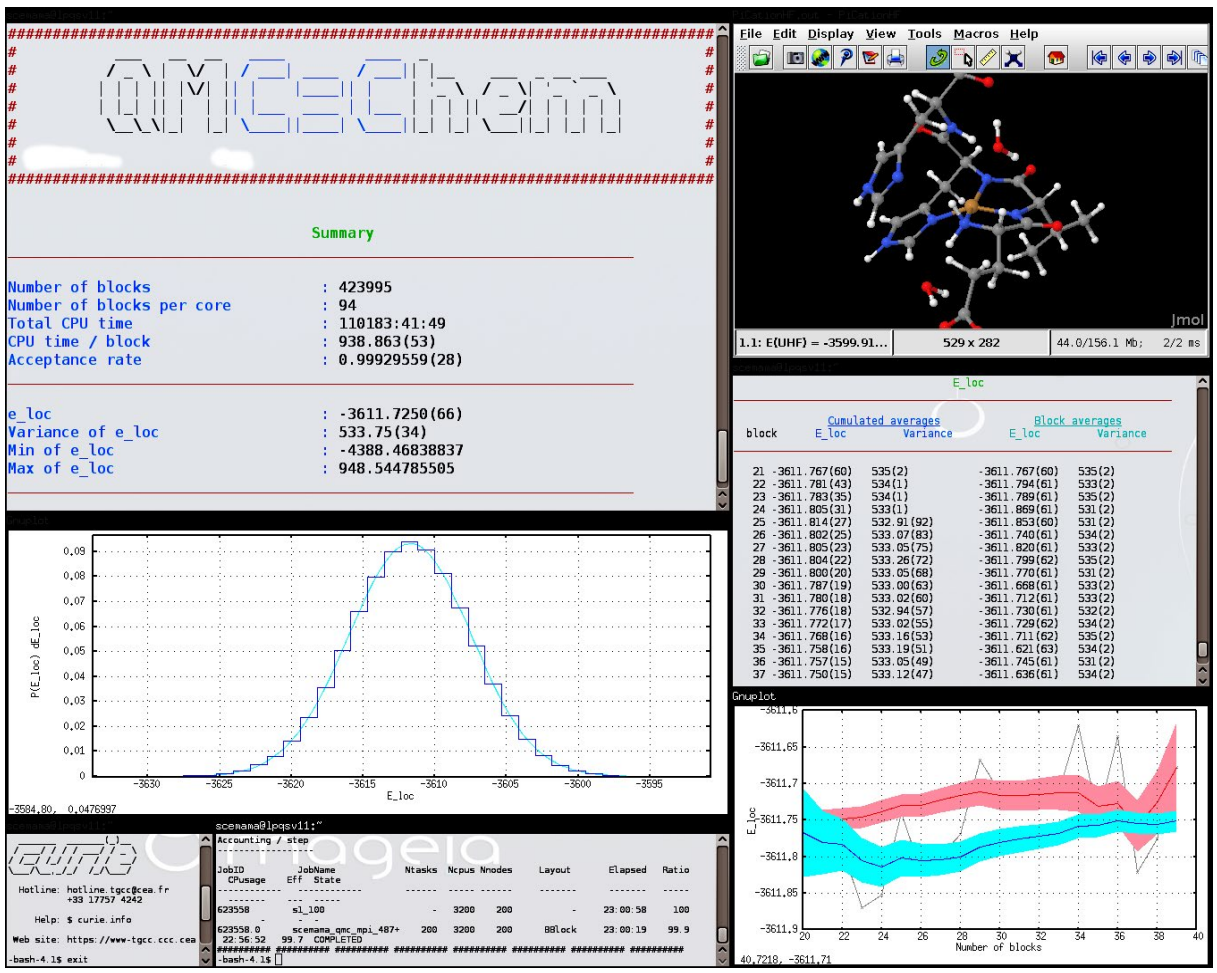


Fig. 1 - A typical QMC simulation.

In parallel with traditional chemistry in the laboratory, a genuine virtual chemistry is developing, where chemical processes are simulated on the computer starting from microscopic equations of matter. With this new "computer chemistry", scientists seek to reconstruct as faithfully as possible the complex electron exchanges at the origin of chemical bonds, interactions between atoms and finally various desired chemical properties.

This problem is a formidable mathematical and data processing challenge since it brings into play the well-known Schrödinger equation of quantum mechanics, whose required solution - the wave function - is a particularly complex function of the set of positions of the atoms' nuclei and electrons.

During the last fifty years, and always in very close relationship with the development of the hardware and software characteristics of computers, several methods have emerged. From an algorithmic point of view, these are mainly based on iterative schemes for solving very large linear systems requiring vast amounts of calculations with stringent I/O and memory constraints. Unfortunately, because they are based on the processing of very large matrices, these approaches do not support massively parallel computation very well.

Our group is developing an alternative method for chemistry - a method quite different from

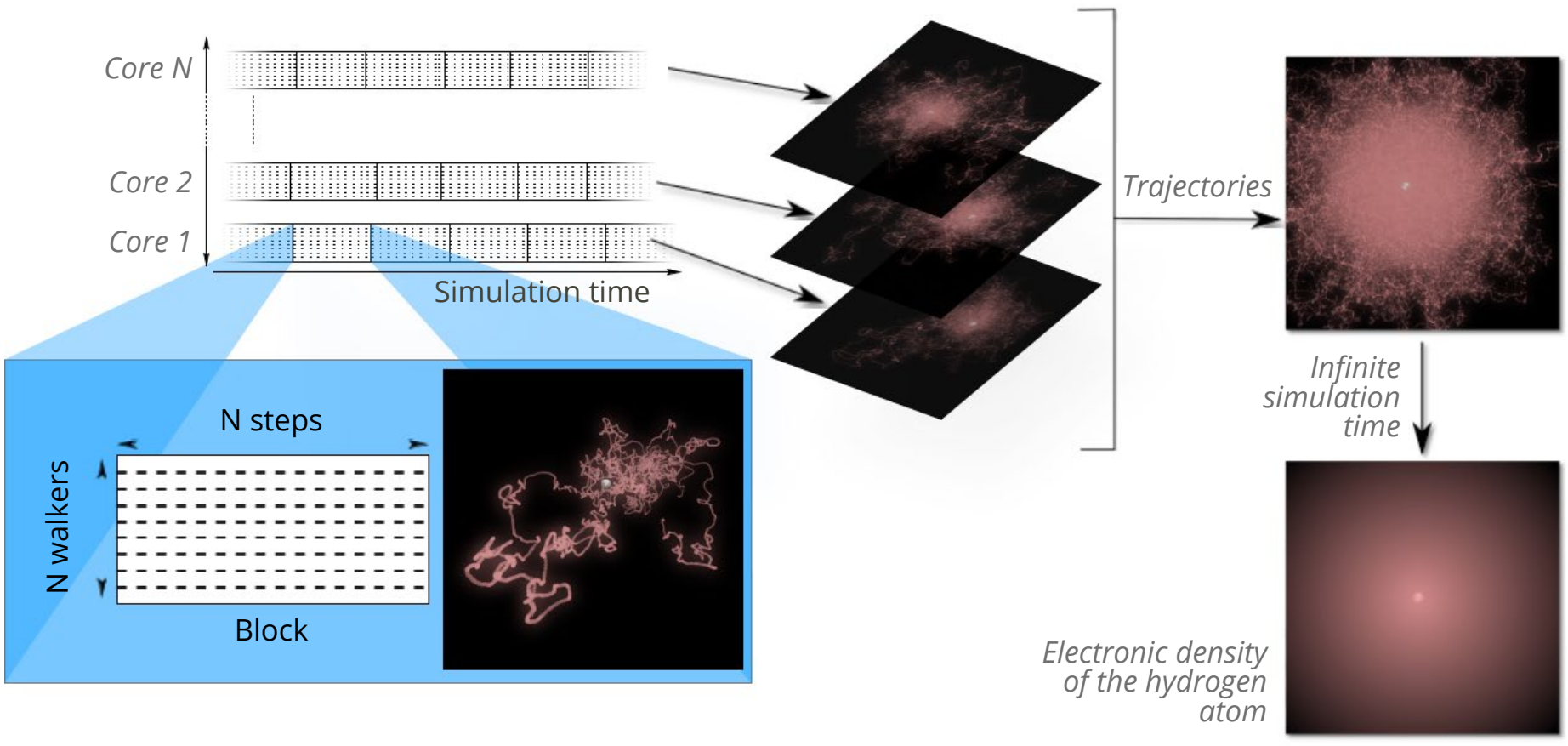


Fig. 2 - Application of the QMC method to the simulation of the hydrogen atom.

the usual techniques but that scales up naturally. Relying on an original interpretation of quantum mechanical probabilities, this "Quantum Monte Carlo" (QMC) approach proposes to simulate the real quantum world, where electrons have a delocalized character, by a virtual world where electrons follow classical trajectories like planets around the sun.

In order to introduce the quantum delocalization absent from this scheme, a random component is added to the movement of the electrons. It is this randomness that gives its name to the method: for each electron displacement, random numbers are drawn in a way similar to the famous Riviera casino roulette. Each trajectory or group of random trajectories can be distributed on an arbitrarily large number of computation cores, each one of these trajectories advancing independently of the others (with no communication between them). The situation is thus ideal from the point of view of parallel computation.

Figure 2 above illustrates the QMC method applied to the case of simulating a simple hydrogen atom (one electron moving around a proton). The image inserted in the blue block represents a few thousands steps of a trajectory of the electron moving randomly around the fixed nucleus (the clearer dot in the figure) of the atom. The trajectories obtained on independent computation cores can be superimposed at will and the exact electron density (probability of presence of the electron) is reconstructed in the limit case of an infinite number of trajectories.

Computational aspects

Efficiently deploying simulations where thousands of electrons carry out billions of step requires that the most critical algorithmic and data processing aspects be optimized as much as possible. So let us now detail our main strategies.

We start by defining a walker as the set of x, y and z coordinates of each one of N electrons of the system (a vector with 3 N dimensions). To accomplish a trajectory, a walker will carry out a random motion in the 3 N dimensional space, in order to reconstruct the probability density corresponding to the square of the wave function.

We also define a block (**Figure 2**) as a set of short trajectories carried out by a group of walkers during a certain number

of steps (about 10,000). For each block, the average values of the properties is computed (energy, dipole moment, etc.) from the successive positions of each walker along all the trajectories of the block. If the blocks are sufficiently long, the final positions of the walkers are completely decorrelated from the initial positions, so the blocks can be considered to be independent. In this case, the average computed values on the blocks have a statistical distribution following a Gaussian law. Each block average is a sample that is used to statistically compute the average values of the properties. The aim is then to compute as many blocks as possible, as fast as possible. For that, we have adopted two strategies. The first one is setting up an effective parallelization system. The second one is reducing the calculation time of each block by working on single core optimization.

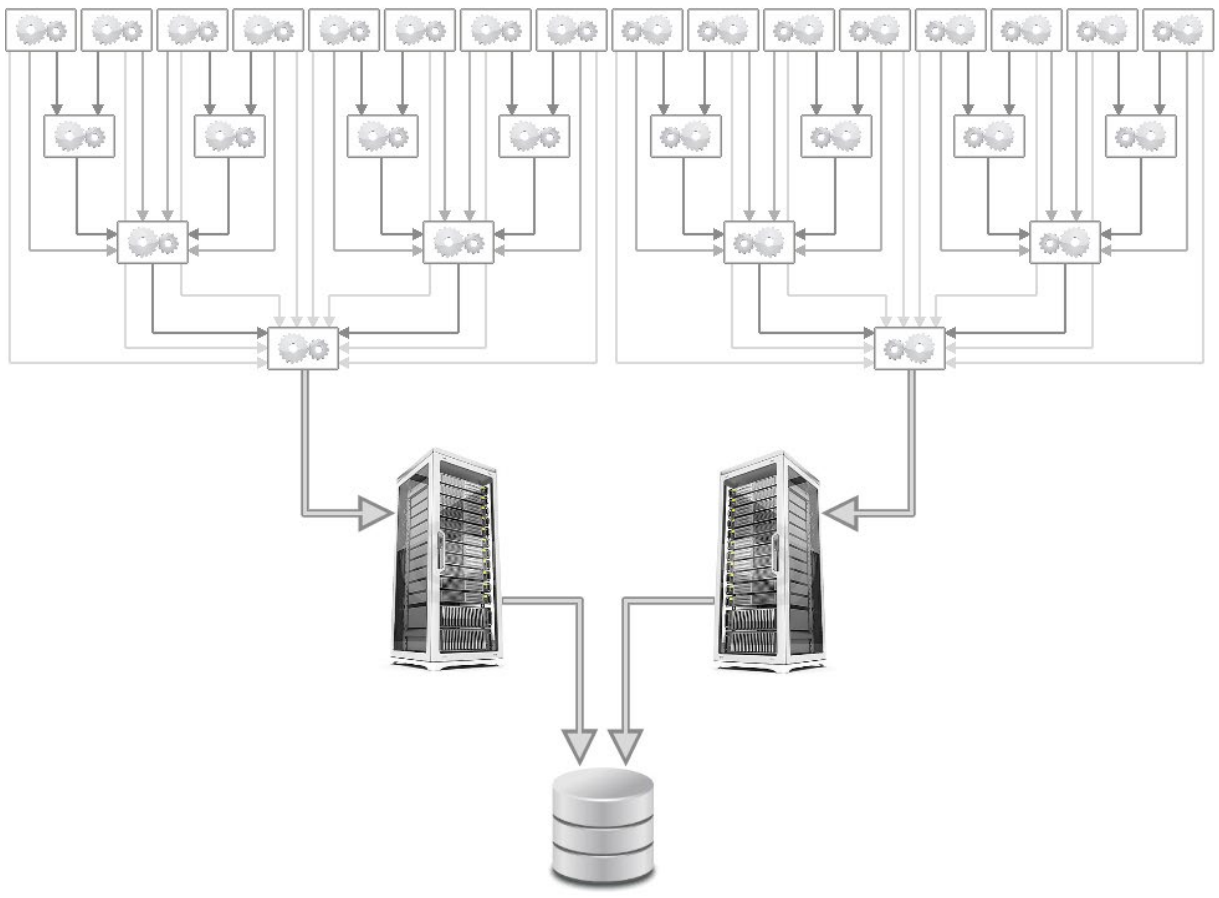


Fig. 3 - Clients are organized as a binary tree to alleviate network communications : a client sends its results buffer to its parent in the tree. When the parent receives data coming from one of its children, it adds them to its own results buffer, which later on will be sent to its parent (the grand parent), etc. The server thus receives a small number of large results packages instead of a large number of small packages. Resilience: if a parent doesn't respond, the client sends its results buffer to the grand parent. If the grand parent doesn't respond either, the client sends its results to the great grand parent and so on, until the packages ends up being sent to the data server.

Managing resilience

With tens of thousands of cores, failure tolerance becomes a new concern. If a computation node breaks down at the end of five years on average, a system with 2,000 nodes has an intrinsic MTBF of about 24 hours. We must then have a system that survives no matter what.

In deterministic algorithms, the calculation is broken down into a number of tasks, each one of them essential for a correct result to be obtained. In a failure event, some of these tasks won't be carried out and the entire calculation will be lost. MPI libraries are well adapted to deterministic calculations. When one MPI client does not respond any longer, the rest of the simulation is lost - as it should. In our model, losing blocks during simulation does not change the averages: it only affects the result's error bar. We thus have chosen not to use MPI for parallelization. Instead, we have implemented a client/server model where clients calculate blocks and send results to the server, which stores them in a database. If one of the client dies, never mind, the others continue with their calculations.

The clients and servers are multithreaded Python scripts communicating by TCP sockets (**Figure 3**). As communications are nonblocking and not very intensive, TCP layer latencies are not a problem. Note that there is a single client per computation node. Each client launches as many calculation programs as there are physical cores on the node. The calculation program is a monothread Fortran binary connected to the client by a Unix pipe. As soon as results are sent to the client, the calculation program immediately starts working on the following block. When the client is inactive, it sends its results to another client so they can be forwarded to the server.

The server comprises two main threads: a network thread and an I/O thread. Data is received by the network thread. It is then

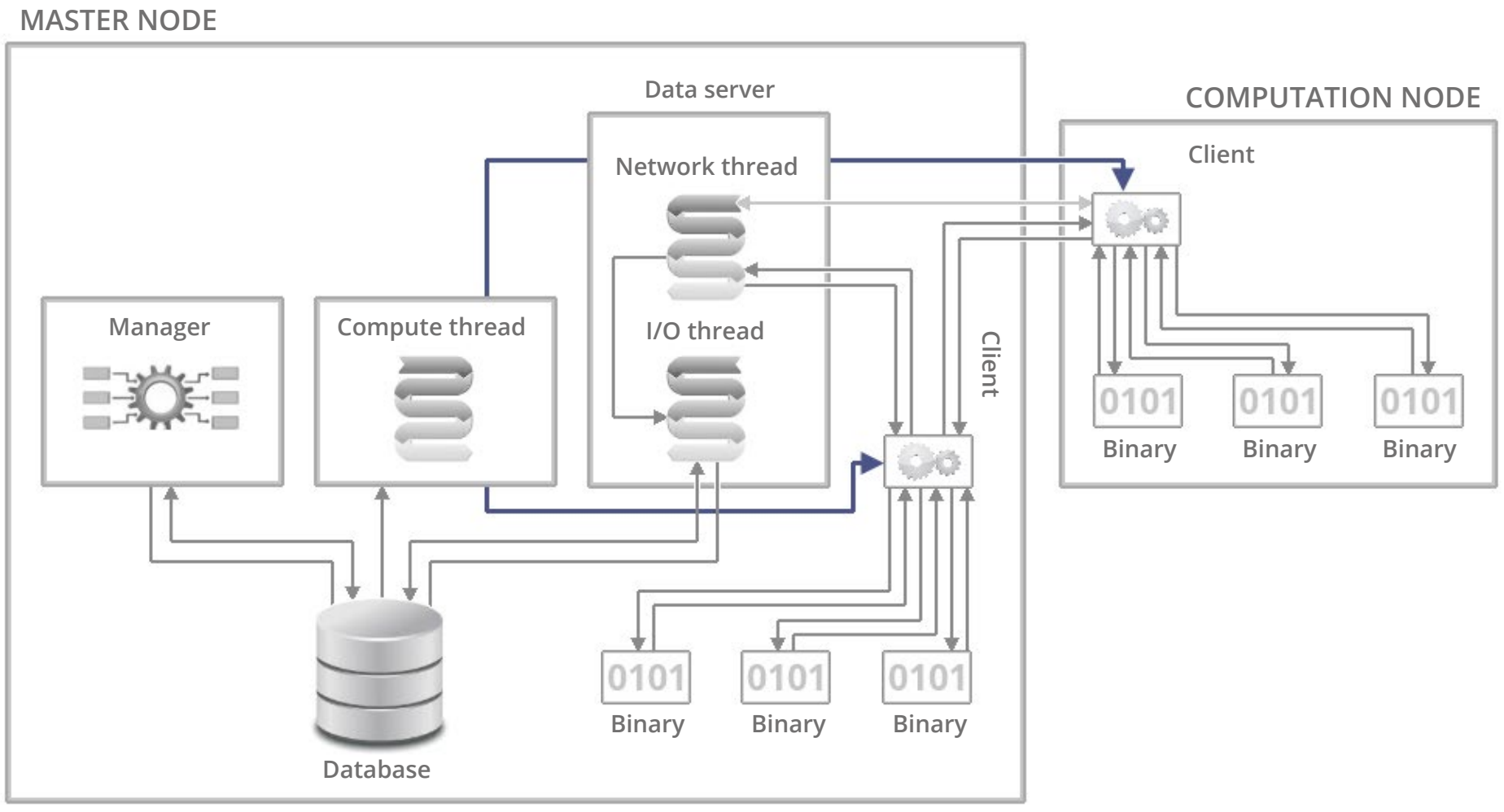


Fig. 4 - Functional diagram of the algorithmic client-server model.

processed and put in a queue. At the same time, the I/O thread empties the queue to save the results to disk. Clients never access the local hard drive or the shared file system but a virtual RAM drive (/dev/shm) to prevent failures connected with temporary storage (disk full, physical breakdown, etc.).

Acceleration factor according to the number of nodes

Since blocks are independent, no synchronization is necessary between clients, which gives an almost ideal acceleration factor according to the number of nodes in the system. However, two critical processing stages can harm the acceleration factor, and must accordingly be dealt with: initialization and termination.

For a fast initialization, it is necessary to start the clients as quickly as possible on the computation nodes. The method that proved the fastest is the use of an MPI launcher that will send by MPI broadcast a tar file containing the Fortran static binary, the Python scripts and the input file on all the nodes. When a node receives the file, it decompresses it and launches the client which immediately starts the calculation programs. When all clients have started, the MPI launcher terminates. Thus, each computation node starts as soon as possible. For a calculation on a thousand nodes, we have measured an initialization time of about twenty seconds.

For termination, taking into account that all the computation nodes are desynchronized, it would be expected that each node has finished calculating its block. This would imply a significant waiting time (many nodes waiting for the last one to finish). We cannot kill all clients either - the results of the blocks in progress would be lost.

We have therefore allowed the calculation programs to intercept the SIGTERM signal in order to cut short the block in progress and to transmit the result of this block to the client. Thus, the termination of a client is almost immediate and no calculation second is lost. For a thousand node calculation, we measured a termination time of about 10 seconds.

All these elements combined yield the curve of **Figure 4**. With a constant number of nodes,

initialization and termination times are constant. In other words, the longer the calculation time, the better the parallel efficiency.

Single core optimizations

Each second gained in the calculation program will have an effect on the totality of the simulation, because no synchronization is blocking. With this in mind, we have undertaken an important single core optimization work on our QMC=Chem program, benefiting from the experience and tools developed by Pr Jalby's group at Intel's Exascale Computing Research Laboratory (Intel-CEA-UVSQ-GENCI) at University of Versailles-Saint-Quentin-En-Yvelines in France.

In our algorithm, at each Monte Carlo step (billions of which are carried out), we must evaluate the wave function, its derivatives with respect to each of the electron coordinates and its Laplacian. These operations use products of small matrices (< 1000 x 1000 elements), of which one is dense and the other is sparse.

The MAQAO statistical analysis tool developed by our Versailles colleagues helped us write a dense matrix x sparse vector routine whose innermost loops theoretically reach 16 flops/cycle in single precision on the Intel Sandy Bridge CPUs of the French CURIE Pflops computer. For that, we have made the following modifications with the purpose of favoring vectorization as much as possible:

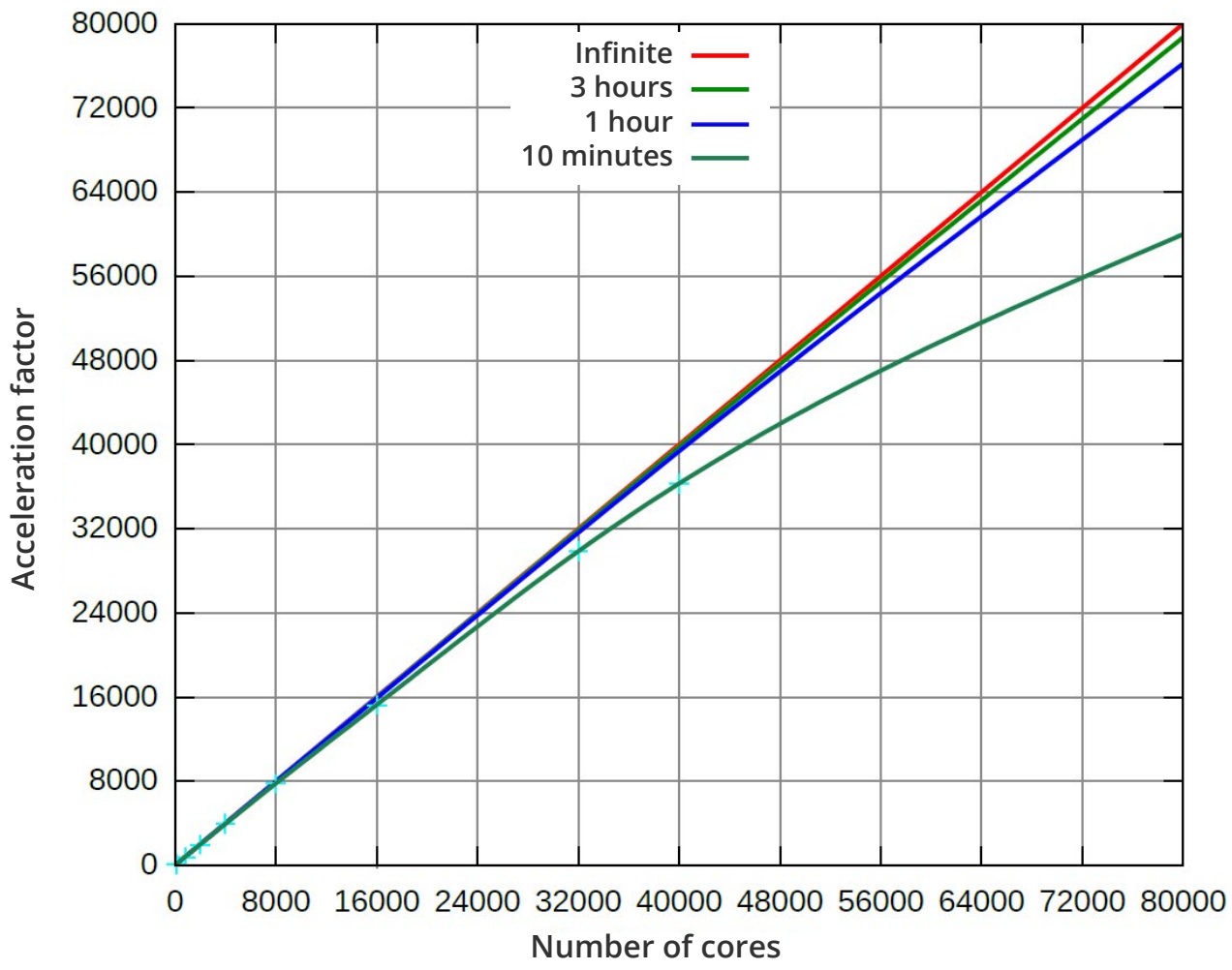


Fig. 5 - Measuring parallel efficiency resulting from the different strategies combined.

• memory accesses are consecutive;

• arrays are 32-bytes aligned, using compiler directives;

• the innermost dimension of the multidimensional arrays is always a multiple of 8 elements, so that each column of the array is 32-bytes aligned;

• the external loop is unrolled (unroll and jam) to reduce the number of memory stores;

• distribution of the loops does not exceed 16 registers, thus reducing accesses to the L1 cache.

In practice, since we do about N^2 operations (N, number of electrons) for N^2 memory accesses, the calculation is inevitably limited by these memory accesses. We have measured up to 61% of peak processor performance in this matrix x vector product routine.

When CURIE was installed in December 2011, we have been able, thanks to the support of the engineers at BULL, the manufacturer of the machine, to carry out our very first large-scale QMC calculation on the approximately 80,000 cores of the machine.

As is often the case with a computer being set up, the calculations underwent interruptions and failures. But what first was a trouble to us eventually proved to be an opportunity: the fact that our simulations went through in spite of these difficulties validated practically the robustness of the client-server scheme.

Petaflops simulations for the chemistry of Alzheimer

Thanks to the exceptional computational resources placed at our disposal these last two years by GENCI (France's Grand Equipement National du Calcul Intensif) and the PRACE European consortium, our group has been able to demonstrate the applicability of our simulations for systems difficult to reach by the usual methods of the field. The applications studied are related to the chemistry of Alzheimer's disease, which brings into play interactions that are particularly tricky to reproduce.

It is known that Alzheimer is associated with a degeneration of the brain connected with the appearance of plaques made up of aggregates of molecules (peptides) known as β-amyloids. Understanding why and how these molecules aggregate is one of the major challenges in understanding the most fundamental mechanisms at the origin of the appearance of this major disease.

In collaboration with the experimental biochemistry group of Pr Faller (LCC, Toulouse), we have been able to show that the chemical accuracy needed to describe these systems quantitatively could be attained by using the CURIE computer. **Figure 5** (on the previous page) illustrates how the random trajectories of the electrons spread out in the case of the simulated chemical systems. During these simulations, we could measure a real sustained performance of about 1 Pflops (960 Tflops simple and double precision to be exact).

Even if much remains to be done, our first probabilistic simulations on a massively parallel architecture open promising perspectives. Although the necessary computational loads - millions of CPU hours typically - are still too important to make QMC a routine method, the forthcoming exaflops machines will probably transform these emerging techniques into a simulation tool accessible to the broadest scientific community. ■

Intelligent Rack PDUs

High Power and Intelligence precisely designed for HPC racks.

Choose from a broad portfolio of high power Intelligent PDUs:

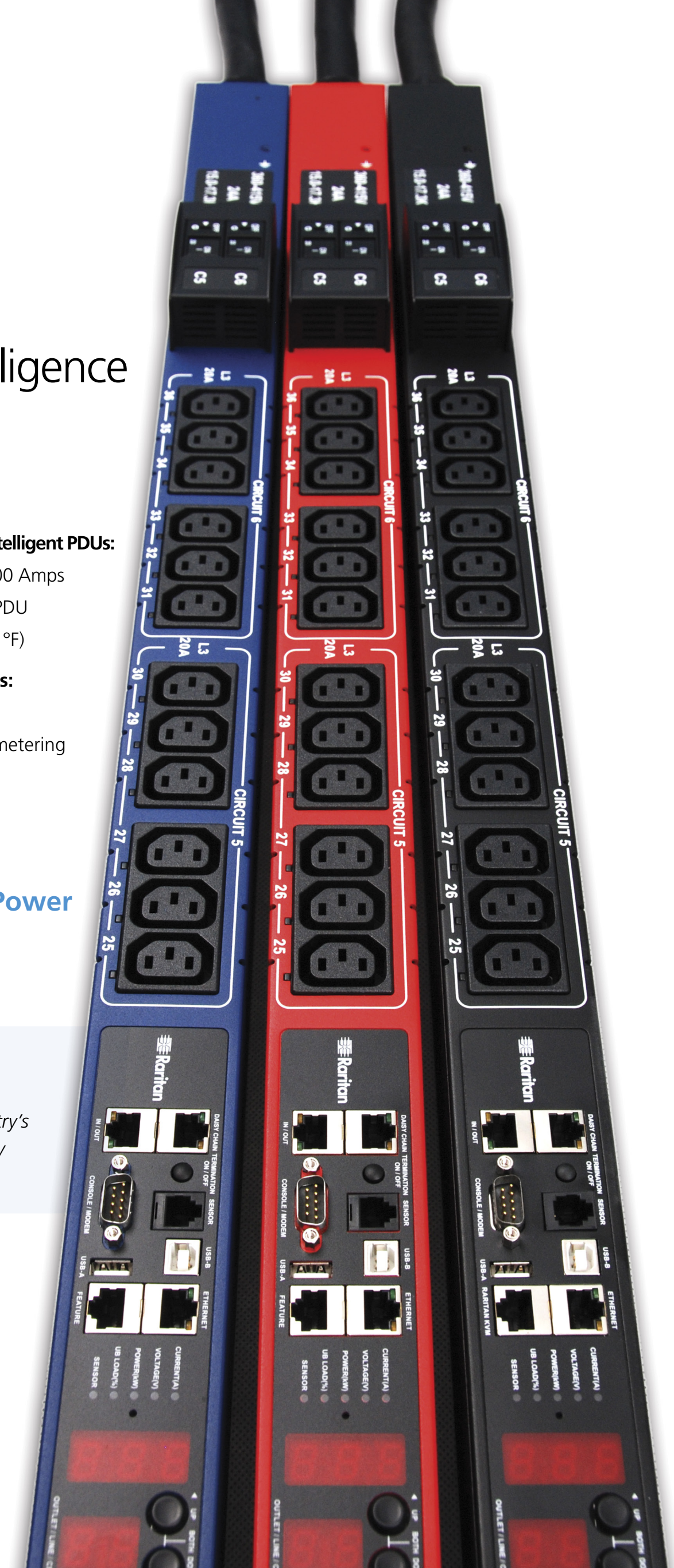
- High Power, capacities up to 55 kW and 100 Amps
- High Density, up to 54 outlets in a single PDU
- Highest ambient temperature (60 °C, 140 °F)

Leverage the industry's smartest capabilities:

- Plug-and-play environmental sensors
- Accurate unit-level and outlet-level kWH metering
- Wi-Fi or wired networking
- Circuit breaker metering and monitoring
- Customizable to fit your HPC racks

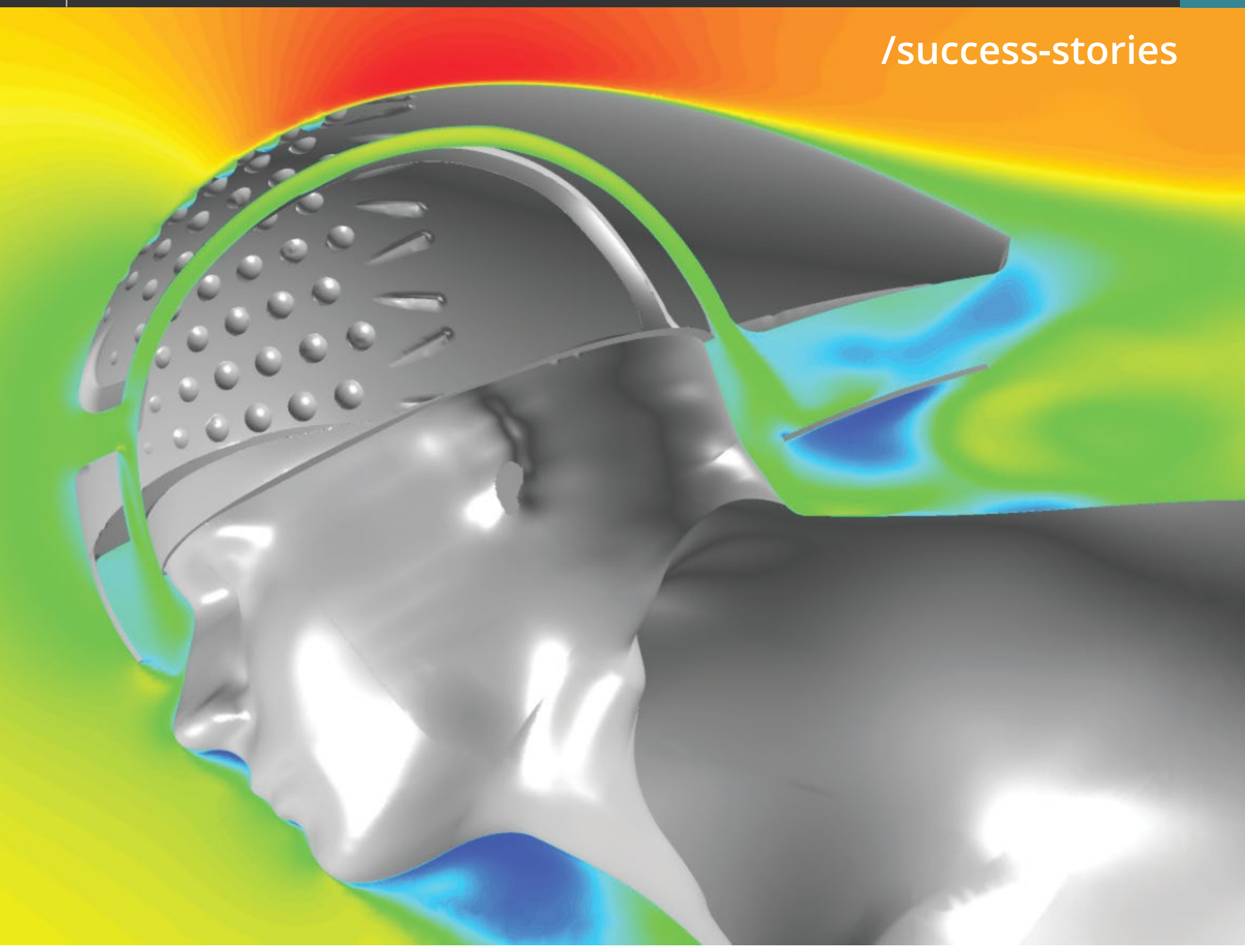
Visit www.raritan.com/SmartPower to learn more and explore all your PDU options.

Blue, red, green, yellow, orange…
and a whole lot more.
Smart managers can choose from the industry's most extensive PDU color palettes to simplify visual identification in their data centers.



Raritan
Know more. Manage smarter.™

/success-stories

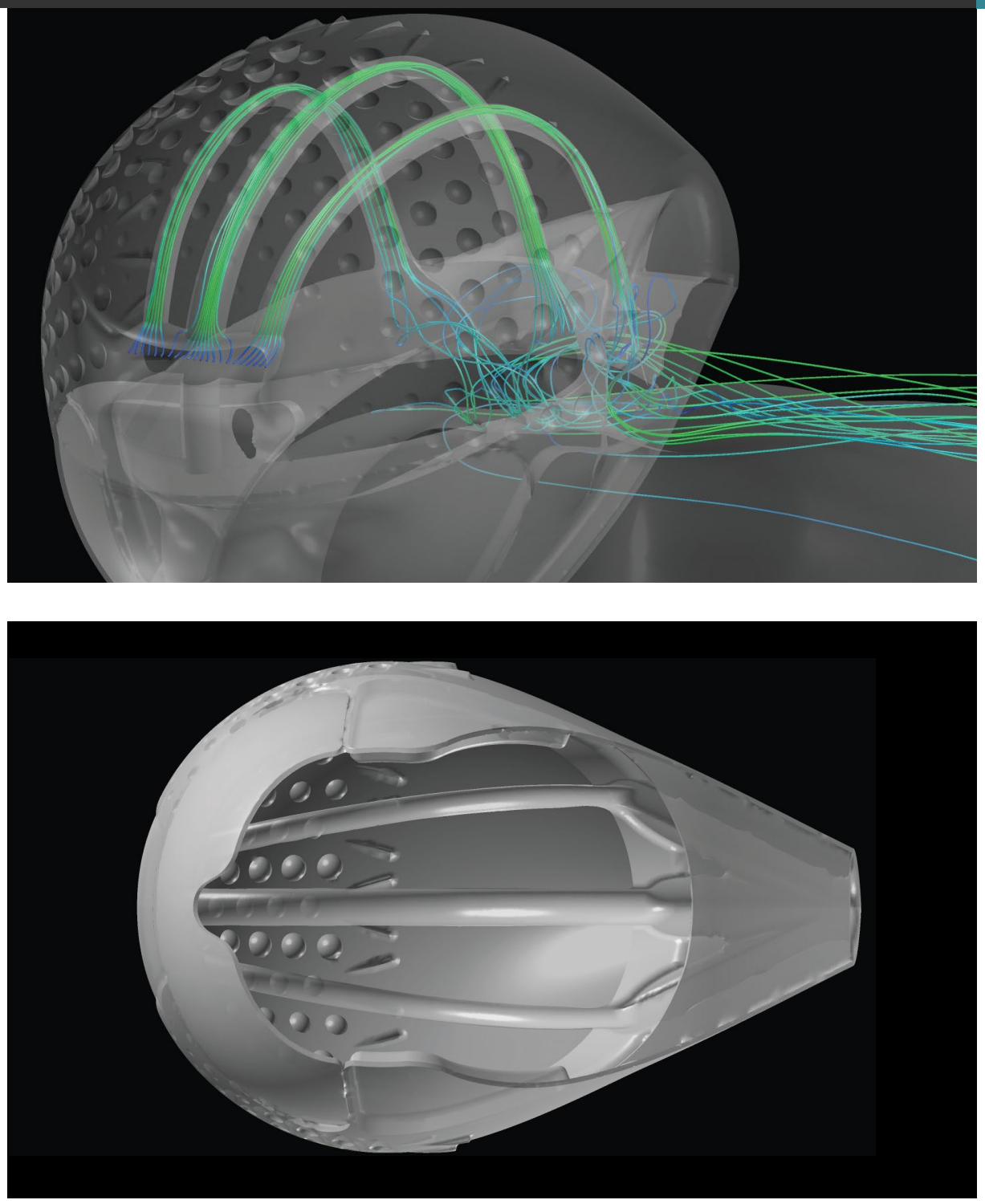


THE LOUIS GARNEAU VORTTICE HELMET: A VICTORY FOR AERODYNAMICS

The latest competition cycle helmet from Louis Garneau, a specialist in top-flight sports equipment, was designed entirely using engineering simulation. The results are convincing, both in racing and financial terms...

Perhaps more than any other sport, competitive cycling is dominated by aerodynamics, as cyclists, clad in aerodynamic clothing and helmets, are forced to adopt uncomfortable crouched position on their bikes to minimize their frontal area and reduce their exposure to the oncoming air. This is most apparent in Individual Time Trial (ITT), which is a key component of triathlon, track and road races. With no other riders to draft behind, the ITT is known as "the race of truth", a brutal contest of man and machine against the clock. Consequently, for around 25 years, professionals have worn specially-designed helmets in order to guide the air around the head and prevent gross separation behind what would otherwise be a bluff body – giving real gains in speed for a given power output.

Year on year, cycling helmets have been the subject of much engineering attention, with the aim of further reducing the drag signature of the rider. To be successful, an aero-helmet also has to be properly ventilated (preventing the rider from overheating at maximum effort), impact resistant (protecting the rider's head in the event of a crash), and constructed from lightweight materials. Determined to address these inherently complex problems as a whole and as innovatively as possible, Louis Garneau turned to engineering simulation, which now plays a leading role in all the company's design and manufacturing processes.



Figs. 1 & 2 - Airflow across the helmet when the rider is in motion and a CAD view of the ventilation channels.

Front runner

Louis Garneau is one of the leading quality sports equipment manufacturers. Its founder and CEO – after whom the company is named – was himself a cycling champion, with over 150 victories over thirteen years, in both road and track events. He represented Canada at the Los Angeles Olympics in 1984.

The first Louis Garneau branded helmet, Prologue, appeared in 2002. It was a revolutionary product, combining aerodynamic streamlining with impact protection. Until then, time trial helmets were basically just head-mounted fairings that offered no protection in the event of a fall. Thanks to its innovative features, Prologue was the first aero-helmet to attain certification to the US CPSC (Consumer Product Safety Commission) bicycle helmet standard, meaning that it was also the first aero-helmet that could legally be sold to the general public. Two years after the arrival of Prologue, cycling's governing body, the UCI, introduced a law stating that all helmets used in competition must meet the less stringent European EN 1078 bicycle helmet safety standard.

From traditional to digital

Through four subsequent iterations (Prologue in 2004, Rocket in 2006, Rocket Air in 2007, and Superleggera in 2008), Louis Garneau continued to advance their professional helmet line, with each new helmet design painstakingly engineered using a traditional design-build-test approach. For each design iteration, it was necessary to create a clay model, tool the production line to make a prototype and perform wind tunnel tests before taking the helmet to market. Making changes after prototyping was difficult, due to the cost and time required to retool the production line. In a competition context, time is especially important, considering that the helmets must be manufactured and shipped in time for set racing days. Although

it was the only practical solution at the time, wind tunnel testing was an imperfect way to measure air flow, especially inside the helmet where air channels are cut to help cool a cyclist's head and around small features such as the dimples on the Superleggera design, whose exact effects were very much unknown.

To get a better understanding of the detailed aerodynamics of their helmets, Louis Garneau approached Lx Research & Development, a mechanical engineering consulting firm that specializes in engineering simulation and product development. To begin integrating CFD (computational fluid dynamics) into the design process, and to demonstrate the concept to Louis Garneau, Lx Research & Development began by scanning a 3-D model of Superleggera and a cyclist to analyze. This required separate scans, which were then merged together digitally to reproduce Louis Garneau's wind tunnel model. The resulting CFD analysis of this model, using the STAR-CCM+ software package developed by CD-adapco, came within 4% of Louis Garneau's own analysis during the wind tunnel test, thus verifying the results found by Lx Research & Development. This was enough to ensure that Louis Garneau engineers adopted the proposed methodology.

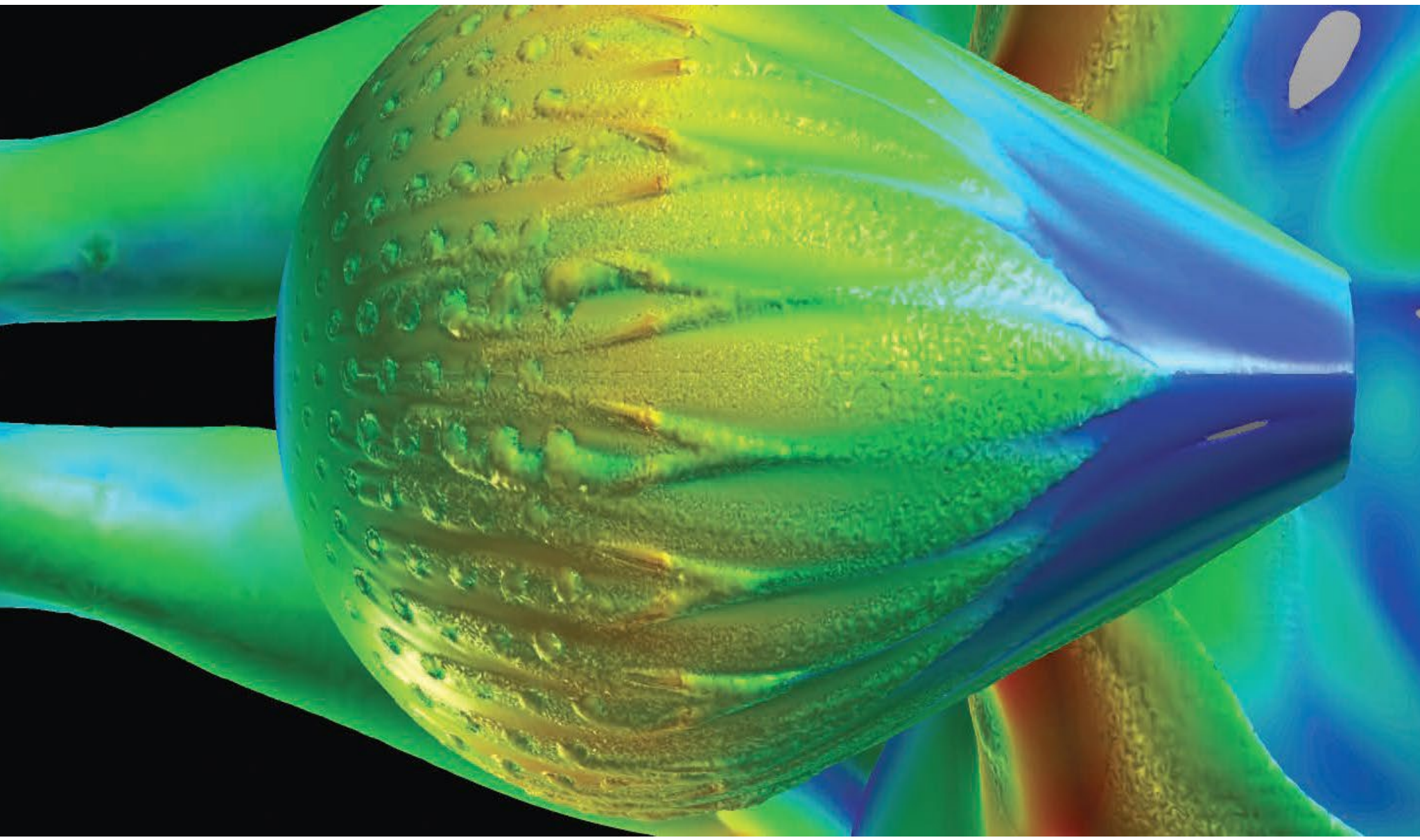


Fig. 3 - Turbulence visualization on the helmet. Note the turbulence at each of the vortex generators.

Winning versions

Boosted by this initial success, the two partners undertook to design a new helmet, Vorttice. In analyzing Superleggera, many ideas were proposed for improvement, such as reducing the frontal area of the helmet, relocating the air intake position, adding air ducts within the helmet to improve air flow and cooling power, and finding ways to reduce the sensitivity of the helmet to angle of attack. A particular idea Lx Research & Development brought to the table was adding vortex generators to the helmet, common in other racing sports, which create turbulent flow across the back half of the hel-

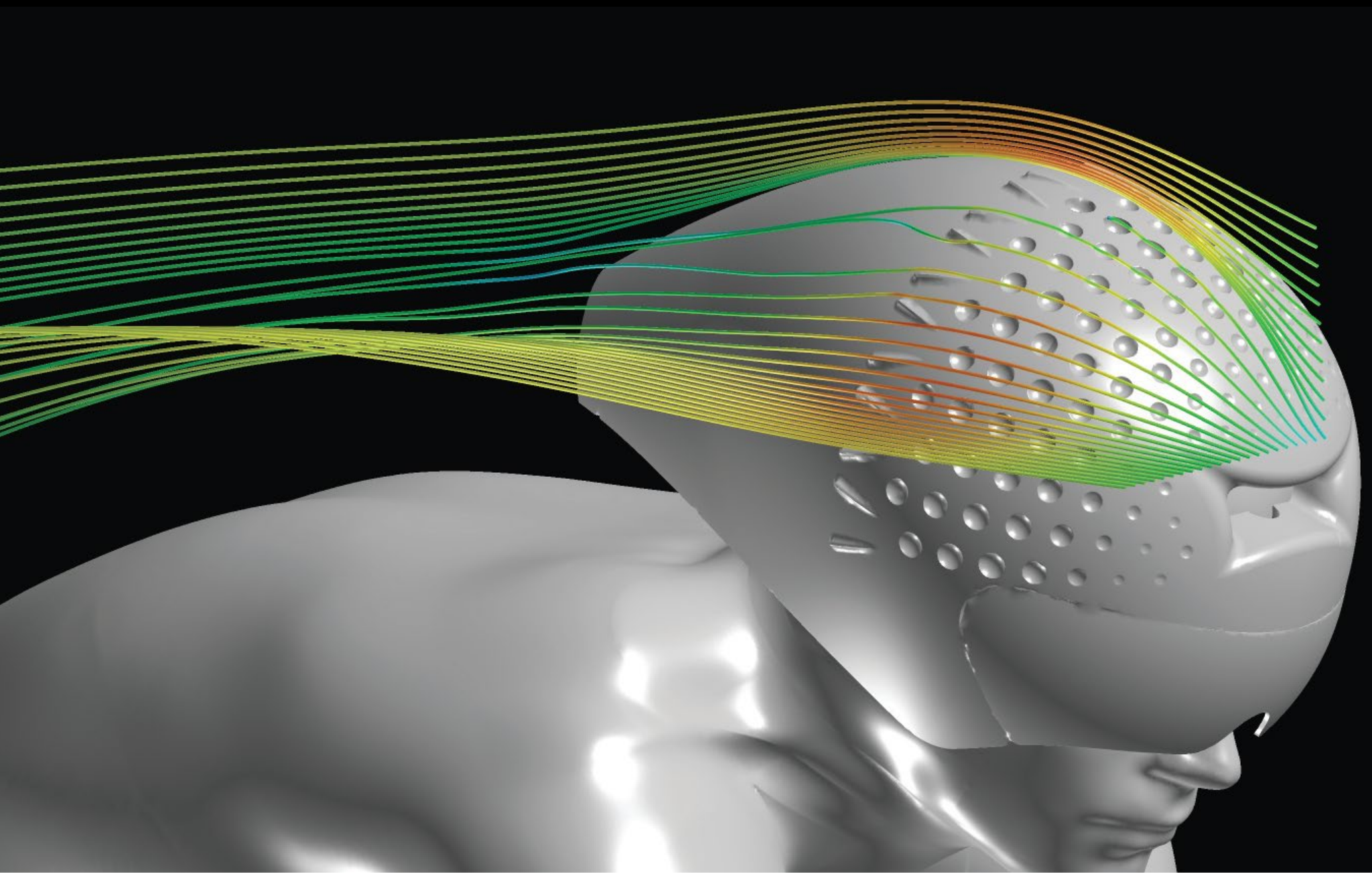


Fig. 4 - Modeling airflow around the helmet.

met, forcing the boundary layer (the interface zone between the body and the fluid) to remain attached for longer and thereby minimizing drag. The two companies were able to make modifications daily to the helmet design, taking information from previous iterations to make small changes.

CFD analyses also highlighted significantly substandard airflow within Superleggera. This led to the decision to alter the location of the cooling channels air intake inside the helmet, which significantly improved its cooling efficiency. Other things that the two companies were able to do were to quickly analyze the helmet's performance at speeds up to 60 km/h and to examine different angles of attack for the helmet, something difficult to reproduce in a wind tunnel because at this speed, a cyclist usually cannot hold his head still enough for accurate measurements to be taken.

Lx Research & Development ran a new helmet design every day (including weekends) for a month before the design was optimized to Louis Garneau's satisfaction. The process produced a vastly superior helmet in a lot shorter period of time than the clay model method, not to mention the reduced cost of the process. In addition, Louis Garneau had vastly more data and information about their new helmet from the CFD analysis than they could have gathered from a wind tunnel test, giving them further insight into potential improvements in the future. After all this, it is hardly surprising that Vorttice became a global success. It was used by team Europcar for the last Tour de France, and also chosen by Mirinda Carfrae, the current Ironman triathlon world champion.

Behind the laboratory doors

Despite the complexity of the design, the final product's CFD analysis uses only 10 million polyhedral cells in the volume mesh for both the helmet and the rider, which significantly reduces the needs in computing power. The re-envisioned process for the design of the hel-

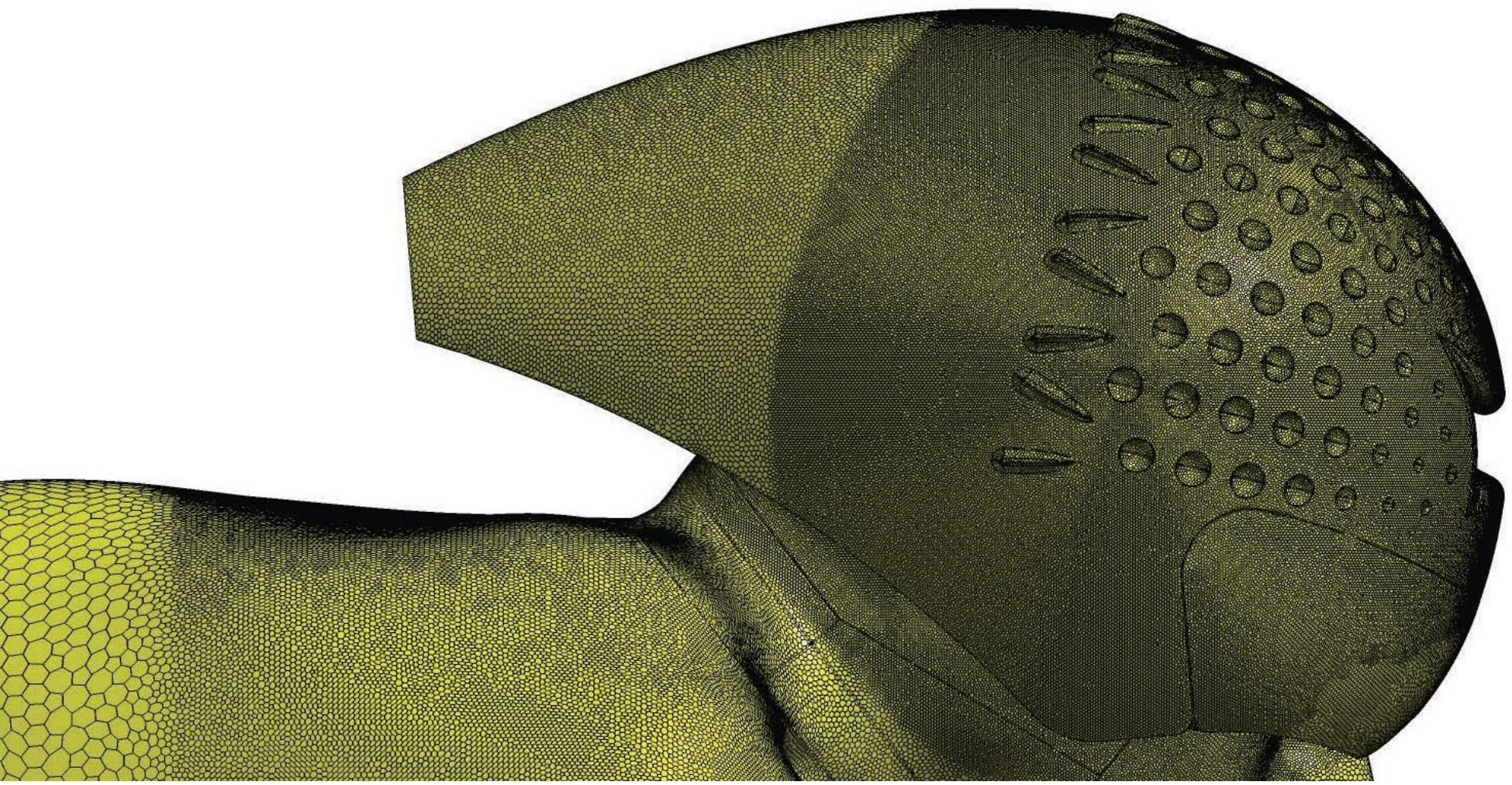


Fig. 5 - Surface mesh of the helmet and rider.

met goes from 3D-CAD modeling, CFD/CAD iterations, rapid prototyping and a single wind tunnel test... to tooling for production. As mentioned above, Lx R&D started the new iteration of the design based simply on a new CAD model. They then wrapped it with the digital cyclist to remove any CAD impurities (using STAR-CCM+'s surface wrapping tool), remeshed and produced results in a single day, reusing only the physics characteristics for each run. These results were then returned to Louis Garneau engineers for analysis and further modifications.

This is a further illustration of the benefits of engineering simulation, compared to the "old-school" process. The gains achieved in technical performance, cost and time have also encouraged Louis Garneau to continue the partnership with Lx R&D. A successor to Vorttice is now being designed, on a brand-new digital engineering platform built around the CD-adapco tools. This may be important news for those of our readers tempted by sporting wagers... ■

Within a few hairs' breadths...

Cycling is a sport in which every second really does count, and small aerodynamic advantages can be the difference between winning and losing the Tour de France. In 1989, Greg Lemond trailed French rider Laurent Fignon by 50 seconds prior to the final stage, a 24.5 km ITT. To most observers, this gap was insurmountable: Lemond had to ride each kilometer two seconds faster than Fignon, himself no mean time-trialist. On a warm Paris afternoon, Lemond, using an aerodynamically streamlined helmet and bicycle, beat Fignon, riding a conventional road bike, by 58 seconds, and won the Tour by just eight seconds. Subsequent analysis suggested that the drag on Fignon's ponytail alone was enough to slow him down by the critical eight seconds, allowing Lemond's name to be on the list of Tour winners forever.

ChilledDoor®
HIGH DENSITY RACK COOLING

NEW

Saves up to 90% power over traditional data center cooling systems

45 kW per rack with 65°F chilled water

- Removes 100% server heat at its source
- No condensate pumps or piping
- Replaces rear door of any server rack
- Hinged door with coil, EC fans & PLC
- Fail-safe Leak Prevention System (LPS)
- Optional Cooling Distribution unit (CDU)
- Quick connect hoses
- Made in the USA

Contact Motivair™ for details

Scan with a QR reader on your smartphone to find out more about the ChilledDoor Rack Cooling System

motivair™
COOLING SOLUTIONS

85 Woodridge Drive | Amherst, NY 14228 | 716-691-9222
info@motivaircorp.com | **www.motivaircorp.com**

CARRI Systems

Technology architect since 1992

intel®

Discover our **XLR** solutions **with PCI Express acceleration cards.**

The extreme computing servers from CARRI Systems are now compatible with Intel® Xeon® Phi™ 5110P coprocessor.

HighServer XLR4i

Power your breakthrough innovations with the highly parallel processing of the Intel® Xeon Phi™ coprocessor. We've packed over a teraFLOPS of double-precision peak performance into every chip—the highest parallel performance per watt of any Intel® Xeon® processor.

intel® inside™ XEON PHI™

Intel® Xeon® Phi™ Coprocessor

HighServer XLR4i
133548

Intel® Xeon® E5-2620 2 Ghz
32 Go DDR3 1600 MHz ECC REG
2 x 1 TB RAID Edition (7 200 rpm, 64 MB cache)
1 x Intel® Xeon® Phi™ 5110P with 8 GB GDDR5 @ 5.0 GT/s
Integrated video chipset
2 x Gigabit Ethernet
2U rack format
2 x redundant power supply - 1620 W
Linux Centos®
3 years on-site warranty (excluding accessories)
Metropolitan France only

Starting at 5 990 euros HT

For more information about XLR solutions

www.carri.com

intel Technology Provider Platinum 2013

intel inside™ XEON®

intel inside™ XEON PHI™

intel inside™ SOLID-STATE DRIVE

/hpc_labs

KEPLER vs XEON PHI:

OUR BENCHMARK [SOURCE CODE INCLUDED]

As promised, here is our first comparative evaluation of the two prominent parallel accelerators readily available for purchase and deployment. Compute, memory, latency - none of the three fundamental dimensions of their programming has been forgotten. The result? Surprising figures and real differences between the official specs and our measurements...

FLORENT DUGUET, PhD*

By now, almost everyone in the HPC community has had a chance to give NVIDIA's Kepler or Intel's Xeon Phi parallel accelerators a spin. But what are they really worth in the field? To find out, scientifically, we submitted them to the very same applicative challenges. On one side of the ring, a commercial Kepler K20X. On the other side, a preproduction sample - the detail is important - of Phi SE10P. Their official specifications are detailed **Table 1** (next page). Declared rivals in the marketplace, both cards share a similar vocation - that is, accelerate your applications - but they feature significant technical differences. It is in taking them into account that we coded the evaluation procedures proposed below.

Simple precisions regarding the architectures

Intel's and NVIDIA's architectures are very similar yet very different at the same time. Both are based on a many/multi-core grid structure including multiply-add floating-point compute units running at

*CEO, Altimesh

clock frequencies of 735 MHz (K20X) *vs* 1.1 GHz (SE10P). Kepler's cores are accessible via API calls through a driver; Phi hosts a complete Linux OS, but its compute machinery is also accessible through an API. So much for the "black boxes".

Let's now dive a bit deeper into their respective internal logics. Kepler accelerators comprise SMX (Streaming Multiprocessor Extended) units that one could compare to CPU cores. Each SMX has its own cache, instruction dispatching units and memory interface. These SMXes (14 in K20X) hold 192 single-precision floating-point units, each of which can perform a multiply-add in one clock cycle, for a headline peak performance of 3.95 Tflops. On the DP front, the units count is down to 64, with an official peak throughput of 1.31 Tflops.

Work distribution on a Kepler accelerator is organized in 32-entries warps. Warps are logically equal between each other - scheduling up to two instructions per cycle - with potential masking. We may compare them to CPU vector units since current AVX systems have 8 single-precision entries. To orchestrate operations, SMXes include four warp schedulers with two dispatch units each.

An SMX can run several contexts at the same time. This context distribution is flexible to some extent, but it is optimal if instructions are the same (SMXes offer single instruction caches). The maximum number of concurrent "threads" to be run is 2048, which in effect mobilizes 64 warps. Hiding latency for some operations (*e.g.* memory access) requires a maximization of the number of concurrently active warps. Note that the number of registers available is 2 Mbits for each SMX, for an approximate total of 28 Mbits in a K20X. This register space is shared among active warps, which narrows it down to 1024 bits per entry - or 32 registers of 32 bits.

Kepler K20X	
Driver version	311.15 (5.0/cuda 5.0)
Device	K20X
Compute capability	3.5
Number of SMX	14
Frequency	735 MHz
Core count	2688
GDDR Size	6 GBytes
GDDR Frequency	2600 MHz
GDDR Bandwidth (ECC off)	250 GB/s

Xeon Phi SE10P (preproduction)	
OS version	2.6.32-279.el6.x86_64
Driver version	5889-16 (GOLD)
MPSS version	2.1.5889-16
Intel compiler version	13.1.1.163 Build 20130313
Device	SE10P
Flash version	2.1.02.0314
uOS version	2.6.38.8-g9b2c036
Number of active cores	61
Frequency	1090909 kHz
GDDR Size	7936 MBytes
GDDR Frequency	2750 MHz
GDDR Bandwidth (ECC off)	352 GB/s

Table 1 - The accelerators' technical specifications.

Intel's Xeon Phi, on the other hand, is an implementation of the x86 MIC (Many Integrated Core) architecture. This architecture is easier to figure: it holds several independent cores (61 in our setup) featuring 512-bit vector units. Each core is hyper-threaded with up to four threads. Vector operations are structurally very similar to SSE or AVX, but with a significantly broader base and

reach (Intel's Reference Guide is more than 700 pages long). A now well-known example of this is the new gather and scatter operations. Designed to ease vector access to memory, they for instance perform a lookup in a single instruction.

Functional analogies

These structural differences notwithstanding, we must consider some analogies between the two architectures, if only from the standpoint of a programmer with a mission to design a consistent benchmarking suite. These analogies, listed in **Table 2**, should not be seen as a reference for comparison, but rather as a guide for programming insights and in view of the brief technical notes at the end of the article.

Based on both accelerators' metrics, we compare vector units with warps, not cores. Indeed, each warp is largely independent and can be paused at some points, just as a thread (within hyper-threaded Xeon Phi cores) would be. This results in comparable vector unit sizes of 32 versus 16. Frequency is similar, or at least in the same range. Each vector unit entry can execute a fused multiply and add in a single cycle. Both architectures also have special function units (SFU) that allow exponential or trigonometric operations, among others, to be computed with excellent throughput. However, these units and the hardwired features they include are so different structurally that we decided to leave them out of this work.

In sync

Given the specificities of both architectures, we had to adopt an appropriate testing approach. In the real applicative world, memory access and compute operations are often executed in parallel, where hiding memory access latency is traditionally achieved by using several parallel contexts (active warps or hyper-threads). So we measured performance using wall-clock execution times that equal the maximum compute + memory access execution durations. Depending on which resource was the bottleneck, we classified problems as compute-bound (calculation time > memory access time) or memory-bound (memory access time > calculation time).

However, the duality is complicated by a third degree. Since Intel's 80386, memory and CPU are not synchronized anymore, which leads to some "natural" latency in memory access. This explains why so many codes spend more time waiting for data to become available than in compute operations - a problem that also impacts memory bandwidth (see Little's law). In such cases, the algorithms at play are neither memory-bound nor compute-bound. We therefore classified them as latency-bound.

Accordingly, we present three classes of problems - one for each category - in several implementations adapted to each of the two hardware platforms. The first problem is memory-bound: we read a large array of floating-point values and sum them up. The second problem is compute-bound: we compose a function a certain number times for which we know how many floating-point operations will be used. The third problem is latency-bound: we access a small amount of data, process it and store a smaller number of values. This code is neither compute-bound nor memory-bound.

Xeon Phi SE10P	Metrics	Kepler K20X
Core (61)	Core (count)	**SMX** (14)
Thread up to **4**	Contexts per core	**Warp** up to **64**
Up to **244**	Total contexts	Up to **896**
Zmm (16/8)	Vector units per context (SP/DP)	**Core** (32)
1090 MHz	Core frequency	**735 MHz**
1	FMA per cycle per vector unit	**1**

Table 2 - Architectural specifications.

Memory or Compute?

Before going further, let's take a look at the hardware capabilities to assess a theoretical boundary between compute- and memory-bound. The latency-bound configuration most often appears on smaller problem sizes, where latency cannot be hidden by other kinds of processing. **Table 3**, based on the vendors' specs, summarizes these capabilities and the number of FP operations that need be performed for a given problem to enter the compute- or memory-bound area.

What the table shows primarily is that we need 24 to 68 times more floating-point operations than memory operations to be compute-bound. This will come as no surprise to HPC programmers, who know that most problems are generally memory related. After all, adding two vectors into a third one amounts to one compute operation and three memory operations. That is why the Gflops specification should not be the preferred performance metrics when choosing a parallel accelerator. We will propose a similar table with measured performance indicators at the end of our evaluation (see **Table 6** at the end of this article).

THEORETICAL	Kepler K20X	Xeon Phi SE10P	Kepler K20X	Xeon Phi SE10P
	float	float	double	double
GFLOPS	3951	2130	1317	1065
GB/s	250	352	250	352
flops/memop	63.2	24.2	42.1	24.2

Table 3 - Flops / memops ratios necessary to qualify a problem as compute-bound (based on the vendors' specifications).

The memory-bound problem

The pseudo-code in **Listing 1** shows the global idea: we initialize and read an array of hundreds of millions of floats, sum-up values and store them in a smaller array, focusing on the read memory-bandwidth rather than the reduce algorithm details. Data is always read in the best possible way.

Three Kepler K20X implementations are provided:

1 - A naïve access to the data for accumulation, in a coalesced way. This first approach does not imply any modification to an existing code.

2 - We use 128-bits loads, reading floats by groups of 4 and doing this twice per thread. This approach requires both vectorization and alignment of the reading blocks (note that alignment can be tested at runtime for just a small overhead in compute performance).

3 - We use the **__ldg** intrinsic to spare the L2 cache, and take advantage of the texture cache channel. In addition to the modifications in implementation 2, operations are read-only.

The results reported **Table 4.A** show correct performance in the naïve implementation, with a penalty of only 15% (or 24% without ECC) compared to the optimal performance.

Four Xeon Phi SE10P implementations are provided:

1 - A naïve approach where we let the compiler vectorize the code automatically.

Listing 1 - Pseudo code for the memory-bound test

```
/// count is several millions
/// no constraint on chunksize
/// initial value of b can be ignored
public void ReadBandwidth(int count, float[] a, int chunksize, float[] b)
{
   for (int chunk = 0; chunk < count / chunksize; ++chunk)
   {
      for (int k = 0; k < chunksize; ++k)
      {
         b[k] += a[k + chunk * chunksize];
      }
   }
}
```

2 - The same naïve implementation with vectorization explicitly disabled to simulate cases where the compiler cannot vectorize.

3 - We use the **_mm512_i32gather_ps** intrinsic to load data in an unaligned way. This requires vectorization.

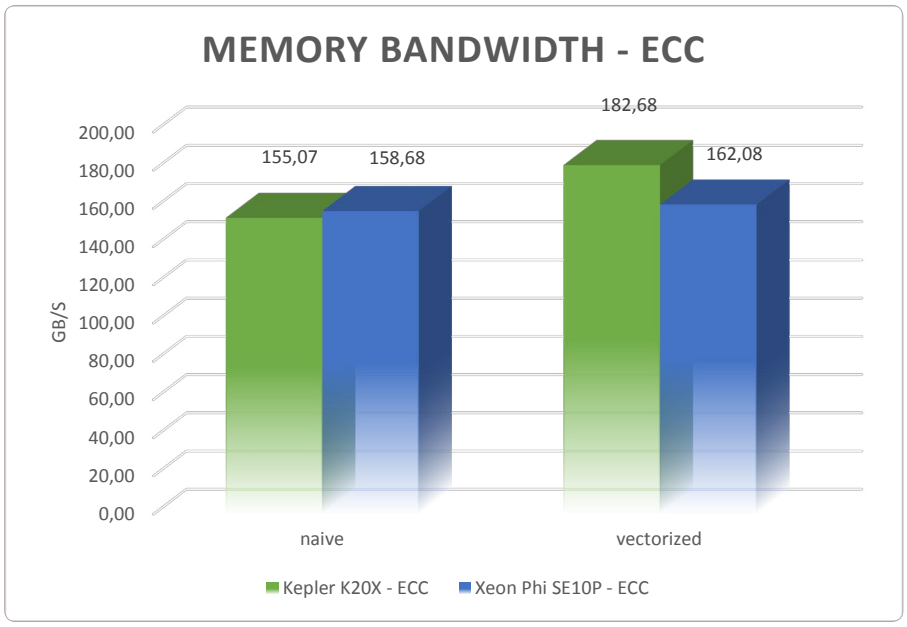
4 - We use the **_mm512_load_ps** intrinsic to load data from aligned addresses.

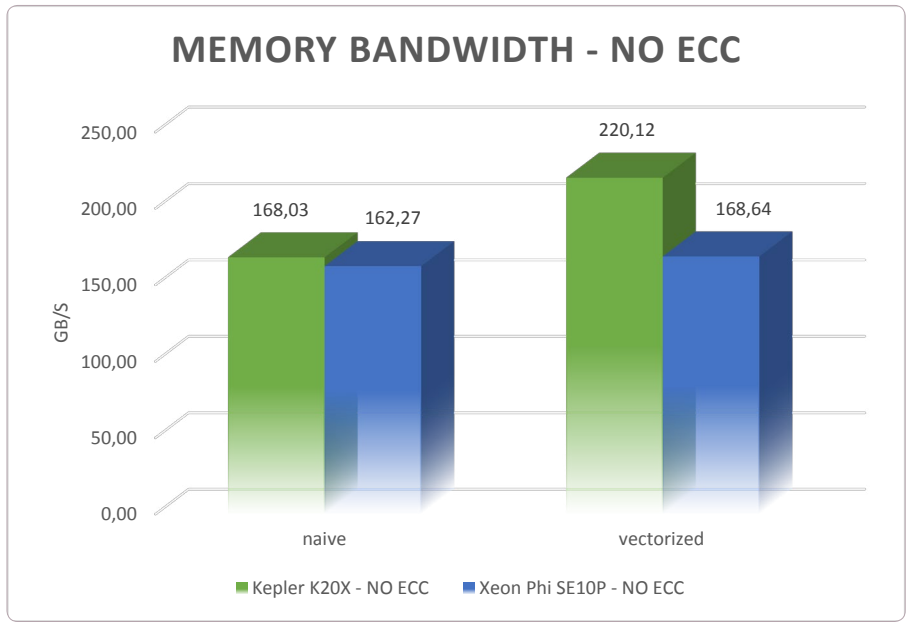
Interestingly enough, the results reported **Table 4.B** show that the naïve implementation with vectorization enabled and hinted gives similar if not better performance than the gather version. From these measurements, it appears the compiler does a fairly good job in this case. Implementation 2, which disables vectorization to see what kind of performance range can be achieved in a more complex example, reveals a performance penalty of (only) 5%.

MEMORY	K20X – ECC	K20X - ECC	K20X – ECC
	Float4	Float4 and __ldg	Float
Achieved GB/s	182,55	182,68	155,07
Theo GB/s	250	250	250
Ratio	73,1 %	73.2 %	62,1 %
	K20X - no ECC	**K20X - no ECC**	**K20X - no ECC**
	Float4	Float4 and __ldg	Float
Achieved GB/s	216,83	220,12	168,03
Theo GB/s	250	250	250
Ratio	86,9 %	88.2 %	67,3 %

MEMORY	SE10P - ECC	SE10P - ECC	SE10P - ECC	SE10P - ECC
	naïve	naïve (no-vec)	gather	load
Achieved GB/s	158.68	142.31	145.43	162.08
Theo GB/s	352	352	352	352
Ratio	45.1 %	40.4 %	41.3 %	46.0 %
	SE10P - no ECC	**SE10P - no ECC**	**SE10P – no ECC**	**SE10P – no ECC**
Achieved GB/s	162,27	148.54	153.70	168.64
Theo GB/s	352	352	352	352
Ratio	46.1 %	42.2 %	43.7 %	47.9 %

Tables 4.A & 4.B - Measured results to the memory-bound test.





The Xeon Phi SE10P preproduction version we used shows results that are globally equivalent to those of our commercial K20X in the naïve implementation. However, if we enforce striding, K20X reveals a significant advantage over Phi, particularly when ECC is disabled.

The compute-bound problem

For this problem, we decided to work with an approximation of the **expm1** function which, as many of our readers probably know, is extensively used in finance for discounting. The function approximation consists of one addition (**add**), two multiplies (**mul**) and five multiply-and-adds (**madd**). We composed the same function twelve times to get stable enough results for positive values below 1/3. This leads to 12 **add**, 24 **mul** and 60 **madd** operations per instruction, which can be safely assumed as being compute-bound. This pseudo-code is shown in **listing 2**.

It must be taken into account that not every algorithm can make full use of the **madd** operation. In this article, we consider **madd** to be one floating-point operation. Most architectures featuring either one-cycle **madd** or same cycle-count **madd** as **add** or **mul**, it is reasonable to view **madd** as a single flop. Moreover, algorithms reconstructing multiply-add instructions based on evaluation graph are well spread among compilers. In that respect, our raw compute power measurements must be viewed as half those of the official specifications.

Two Kepler K20X implementations are provided:

1 - A naïve version that processes floats one by one.

2 - A vectorized version that processes four elements at a time. In both cases, we tested

Listing 2 - Pseudo code for the compute-bound test.

```
public float myfunc(float x)
{
 float res = x + 8;
 res = res * x + 56;
 res = res * x + 336;
 res = res * x + 1680;
 res = res * x + 6720;
 res = res * x + 20160;
 return res * x * 2.48015873015873015873015873015873e-5f;
}

public void Compute(int count, float[] a, float[] b)
{
 for (int k = 0; k < count; ++k)
 {
  b[k] = myfunc(myfunc(myfunc(myfunc(
   myfunc(myfunc(myfunc(myfunc(
   myfunc(myfunc(myfunc(myfunc(
    a[k])))))))))));
 }
}
```

Listing 3 - An example of naïve compute code for Kepler K20X.

```
__global__ void expkernel(int arraySize, const float *  a, float* b)
{
 for (int i =  threadIdx.x + blockDim.x * blockIdx.x ;
    i < arraySize ; i += blockDim.x * gridDim.x)
 {
  b [i] = myexpm1 (myexpm1 (myexpm1 (myexpm1 (myexpm1 (
   myexpm1 (myexpm1 (myexpm1 (myexpm1 (myexpm1 (
   myexpm1 (myexpm1 (a [i])))))))))) ;
 }
}
```

Workloads are dispatched on the processor through initialization and incrementation of the indices. Each thread/block in the execution grid operates on a different element.

Listing 4 - An example of optimized compute code for Kepler K20X.

```
__global__ void expkernelfloat4(int arraySize, const float4 * a, float4 * b)
{
 for (int i = threadIdx.x + blockDim.x * blockIdx.x ;
    i < arraySize ; i += blockDim.x * gridDim.x)
 {
  b [i] = myexpm1float4 (myexpm1float4 (myexpm1float4 (
   myexpm1float4 (myexpm1float4 (myexpm1float4 (myexpm1float4(
   myexpm1float4 (myexpm1float4 (myexpm1float4 (myexpm1float4(
   myexpm1float4 (__ldg(&(a [i])))))))))))) ;
 }
}
```

Overload functions so they can operate on 4 values simultaneously is the only notable modification to the code in Listing 3. Interlacing arithmetic functions gives the hardware a good opportunity to implement instruction-level parallelism (ILP).

performance in single (IEEE-754, 32 bits) and double (64 bits) precision.

The results reported **Table 5.A** show an excellent system usage ratio in double-precision, with or without vectorization. However, single-precision performance without vectorization is disappointing. The reason is simple: SMXes have 192 cores that can be organized in 6 batches of 32 units (each one processing a warp). The accelerator would then reach its highest level of performance when 6 warps execute instructions on the same cycle. But since we only have 4 warp schedulers, they would need to issue two instructions for the same cycle. In other words, using scalar operations in a naïve implementation does not allow the compiler and the hardware to take advantage of the available resources. However, switching to four floats per instruction, we reach a much higher usage ratio (above 66%). These numbers confirm the interest of instruction level parallelism here.

Two Xeon Phi SE10P implementations are provided:

1 - Naïve processing.

2 - Vectorization based on intrinsics and using **m512** registers. In this approach, we use two **m512** registers per entry, *i.e.* 32 float entries per vector. This indeed yields better results, whereas going to four decreases performance. Note that given the in-order context, we also added some cache prefetching instructions to optimize the data feeds.

Listing 5 - An example of naïve compute code for Xeon Phi SE10P.

```
__attribute__((target(mic))) void expkernel(int size, const float* __restrict a,
float* __restrict c)
{
 #ifdef __MIC__

 #pragma omp parallel for
 #pragma simd
 for (int k = 0 ; k < size ; ++k)
 {
  c[k] = myexpm1(myexpm1(myexpm1(myexpm1(myexpm1(
   myexpm1(myexpm1(myexpm1(myexpm1(myexpm1(myexpm1(
   myexpm1(a[k])))))))))) ;
 }

 #else
 #endif
}
```

Using **pragma omp parallel for** and **pragma simd** gives very good results even with a very naïve version of the code.

Listing 6 - An example of optimized compute code for Xeon Phi SE10P.

```
__attribute__((target(mic))) void expkernel_hand_omp(int size, const float*
 __restrict a, float* c)
{
#ifdef __MIC__
#pragma omp parallel
{
 int startK = omp_get_thread_num() * (size/32) / omp_get_max_threads() ;
 int endK = (1 + omp_get_thread_num()) *
   (size/32) / omp_get_max_threads() ;

 #pragma noprefetch
 for (int k = startK ; k < endK ; ++k)
 {
  bival ct ;
  ct.x = _mm512_load_ps ((void*)(a + 32 * k)) ;
  ct.y = _mm512_load_ps ((void*)(a + 32 * k + 16)) ;
  ct = myexpm1V(myexpm1V(myexpm1V(myexpm1V(myexpm1V(
     myexpm1V(myexpm1V(myexpm1V(myexpm1V(myexpm1V(
     myexpm1V(myexpm1V(ct))))))))))) ;
  _mm512_storenrngo_ps ((void*)(c + 32 * k), ct.x) ;
  _mm512_storenrngo_ps ((void*)(c + 32 * k + 16), ct.y) ;
  _mm_prefetch ((const char*)&a[32*(k+1)], _MM_HINT_T0) ;
  _mm_prefetch ((const char*)&a[32*(k+1)+16], _MM_HINT_T0) ;
  _mm_prefetch ((const char*)&a[32*(k+8)], _MM_HINT_T1) ;
  _mm_prefetch ((const char*)&a[32*(k+8)+16], _MM_HINT_T1) ;
 }

}
#else
#endif
}
```

Note the use of software prefetch through **pragma noprefetch** and the **_mm_prefetch** functions.

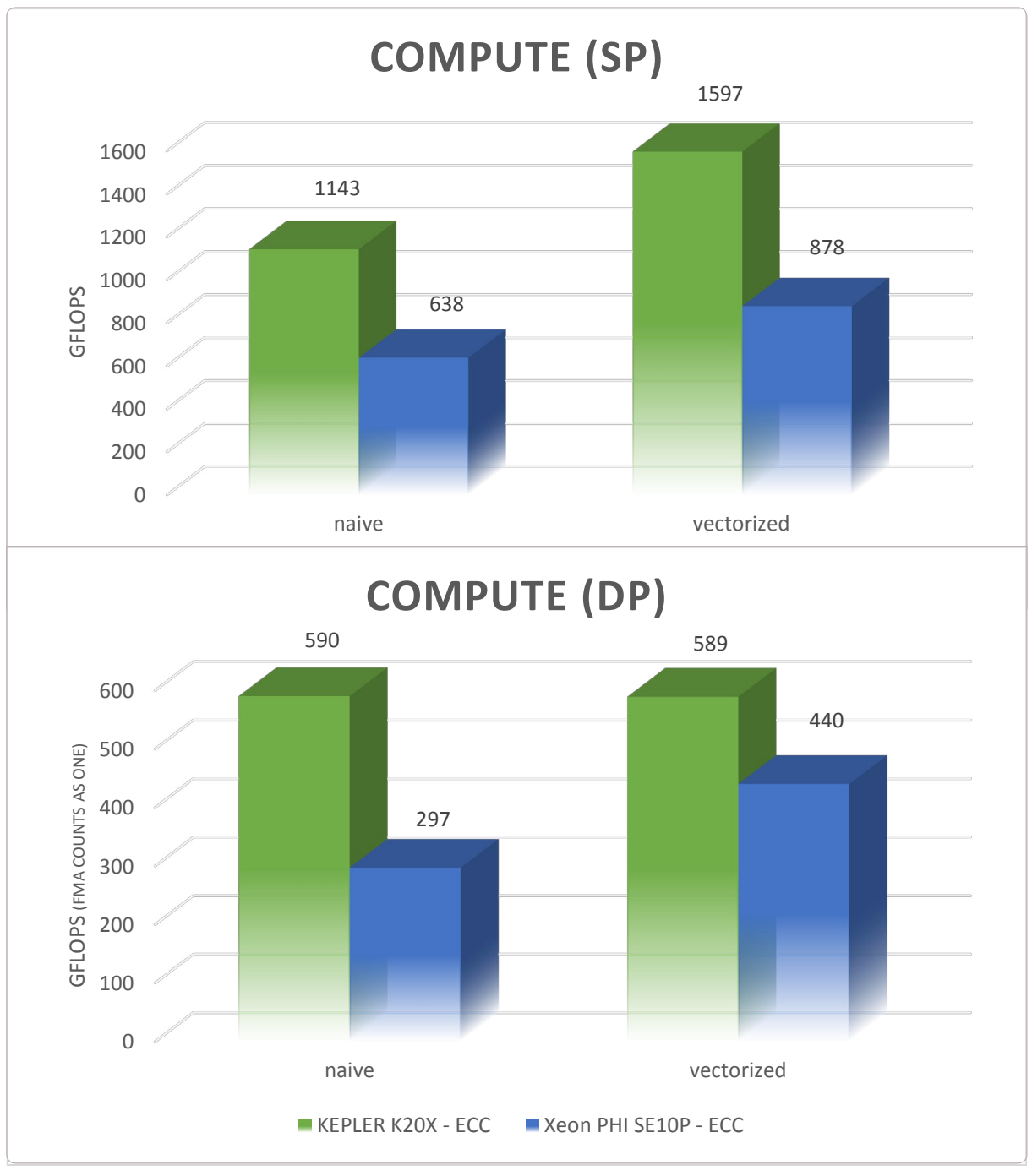
COMPUTE	K20X - ECC	K20X - ECC	K20X - ECC	K20X - ECC
	Float4	Double4	Float	Double
Achieved GFLOPS	1597	589	1143	590
Theo GFLOPS	1968	656	1968	656
Ratio	81,1 %	89,8 %	58,1 %	89,9 %
	K20X - no ECC	**K20X - no ECC**	**K20X - no ECC**	**K20X - no ECC**
Achieved GFLOPS	1599	589	1142	591
Theo GFLOPS	1968	656	1968	656
Ratio	81,3 %	89,8 %	58,0 %	90,1 %

COMPUTE	SE10P - no ECC	SE10P - no ECC	SE10P - ECC	SE10P - ECC
	naïve	intrinsics	naïve	intrinsics
	Float	Float	Float	Float
Achieved GFLOPS	640	879	638	878
Theo GFLOPS	1065	1065	1065	1065
Ratio	60,1 %	82,5 %	59,9 %	82,4 %
	Double	Double	Double	Double
Achieved GFLOPS	297	440	295	439
Theo GFLOPS	533	533	533	533
Ratio	55,8 %	82,6 %	55,4 %	82,4 %

Tables 5.A & 5.B - Measured results to the compute-bound test.

Numbers in **Table 5.B** show that performance improves when we ask the compiler to vectorize but remains below what we could obtain with a manual implementation. However, with intrinsics for vectorized code, we can make use of more than 80% of the hardware, a figure in the same range as Kepler's. Last, we remark that the naïve implementation in double-precision is significantly slower than its K20X equivalent.

Looking at the SP graph at right, you can see that the naïve implementation on K20X yields excellent performance, even without ILP, while the vectorized version is totally out of reach by our Phi SE10P sample. In double-precision, Kepler outperforms Xeon Phi both in naïve and vectorized implementations. It must also be noted that Kepler's naïve code is faster than Phi's optimized version. The conclusion is that highly



compute-intensive codes may benefit from GPUs. Now, let's not forget that pure compute-bound algorithms are not so frequent. For instance, composing our function eight times only instead of twelve would have reset the evaluation code as memory-bound.

The latency-bound problem

To evaluate our accelerators in this dimension, we chose to use an accessor to a lookup table, as lookups are not necessarily predictable by compiler and/or hardware, which is exactly what we need in a benchmark context. We kept the number of iterations small, in order to demonstrate latency in the read operations.

Note that the index to the look-up table is the same for any number of entries, leading to a natural vectorization axis. For each product, we used the best possible size given the minimal work distribution. The pseudo-code is given in **Listing 7**.

Three Kepler K20X implementations are provided:

1 - Naïve.

2 - Swapping the two loops.

3 - Using **__ldg** and processing entries by groups of eight.

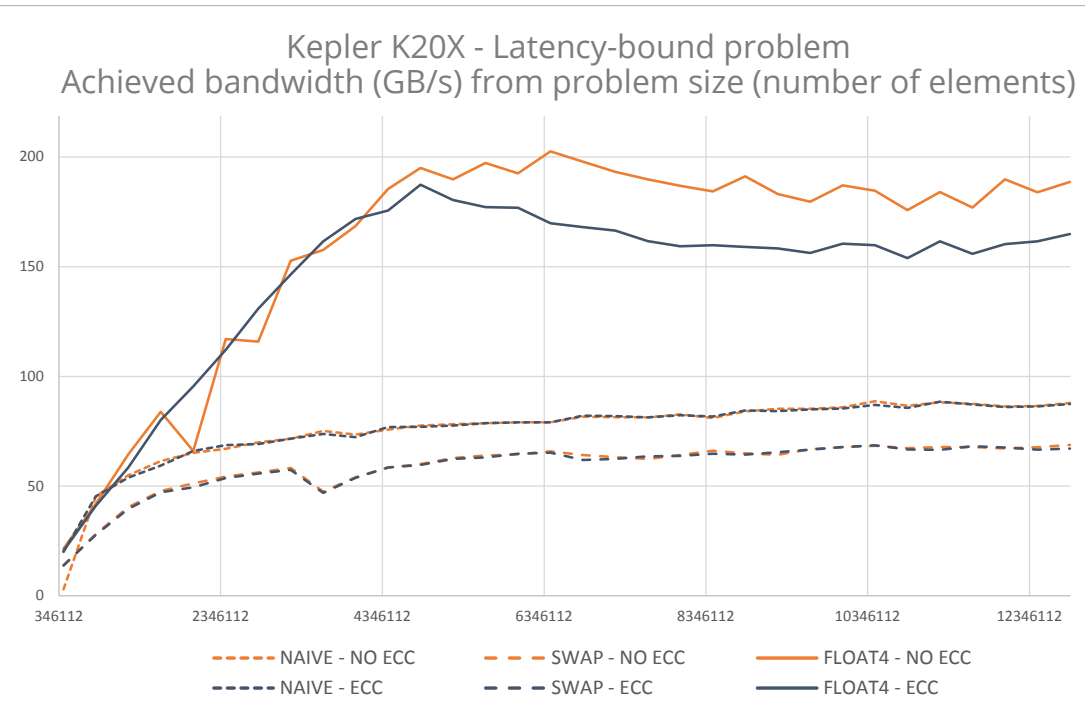
Two Xeon Phi SE10P implementations are provided:

1 - Naïve.

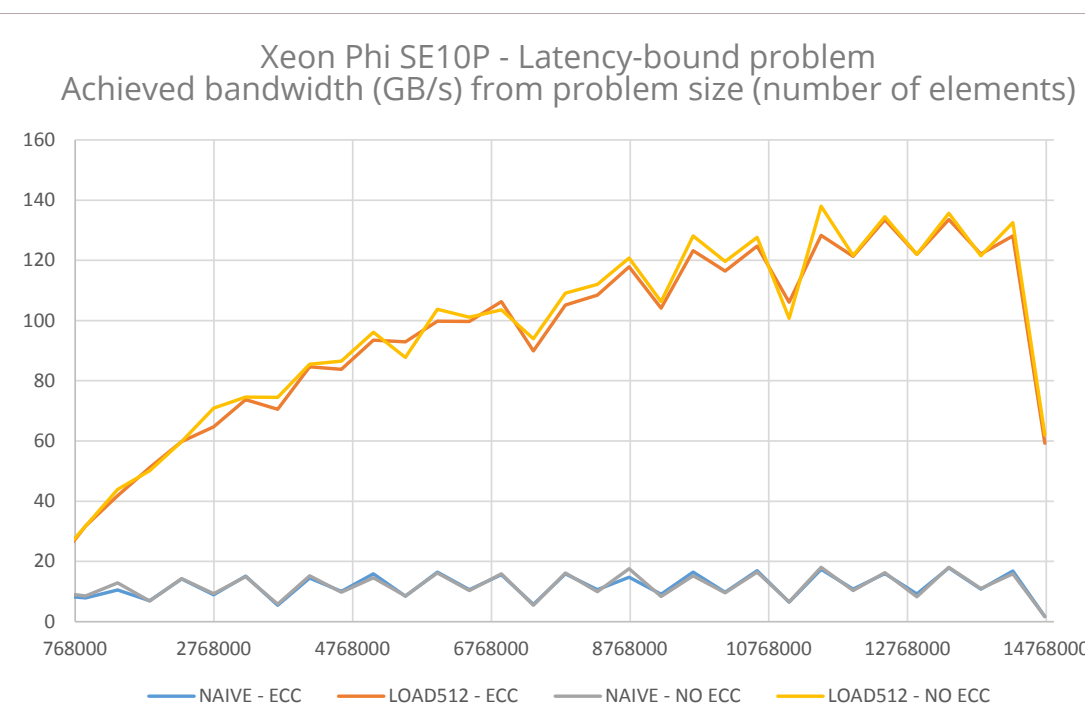
2 - Intrinsic-based, with the loading of a complete 512-bit vector for each iteration.

Listing 7 - Pseudo code for the latency-bound test.

```
/// count is the problem size
/// numidx is 60 in our tests
public void Latency(int count, float[] a, int numidx, int[] index, float[] b)
{
   for (int i = 0; i < count; ++i)
   {
      float tmp = 0;
      for (int k = 0; k < numidx; ++k)
      {
         tmp += a[i + index[k]];
      }
      b[i] = tmp;
   }
}
```



Kepler's SMXes being able to manage up to 8 blocks simultaneously, they can be viewed as hyper-threaded cores with 8 threads by core. That is why we used several problem sizes for our three implementations.



Xeon Phi cores being able to run up to 4 threads simultaneously, we also used several problem sizes according to the hardware specs of the accelerator. Sharp dips can be remarked in the curves, for which we have no explanation.

The graphs above highlight performance per problem size in achieved read bandwidth for the main array. Remark that Phi's naïve implementation barely reaches 20 GB/s, a figure close to what could be reached by plain CPU setups. On the contrary, memory bandwidth in K20X is available even for smaller problem sizes.

Revisiting flops/memops

At the beginning of this article, **Table 3** presented a comparison of performance figures based on the vendors' official specs. The figures induced a flops/memops indicator that could be used to figure out whether an algorithm is compute- or memory-bound. Following our evaluation, it seemed pertinent to propose the same table again (**Table 6**) based on measured results.

These results call for a few comments. First, the bandwidths we achieved are clearly lower than the theoretical ones. Is anyone surprised? Secondly, the advertised Gflops are not necessarily available, and counting fused multiply-adds as single flops seems more relevant when counting flops, since many algorithms do not map trivially on such operations. Last, the memory/compute ratios remain very high (and roughly similar in SP and DP): we need up to 35 floating-point operations per memory operation to be compute-bound. Regarding this particular point, take into account that if transcendental functions are to be used, algorithms might be compute-bound with Xeon Phi and memory-bound with Kepler. ■

Download our complete test code

BEST RESULTS	Kepler K20X	Xeon Phi SE10P	Kepler K20X	Xeon Phi SE10P
	float	float	double	double
GFLOPS	1597 (3951)	878 (2130)	590 (1317)	439 (1065)
GB/s	182,68 (250)	162,08 (352)	182,68 (250)	162,08 (352)
flops/memop	35 (63,2)	22 (24,2)	26 (42,1)	22 (24,2)

Table 6 - Flops / memops ratios necessary to qualify a problem as compute-bound. The numbers are based on measured results and should be compared to those based on the vendors' hardware specifications (between parentheses).

Technical notes

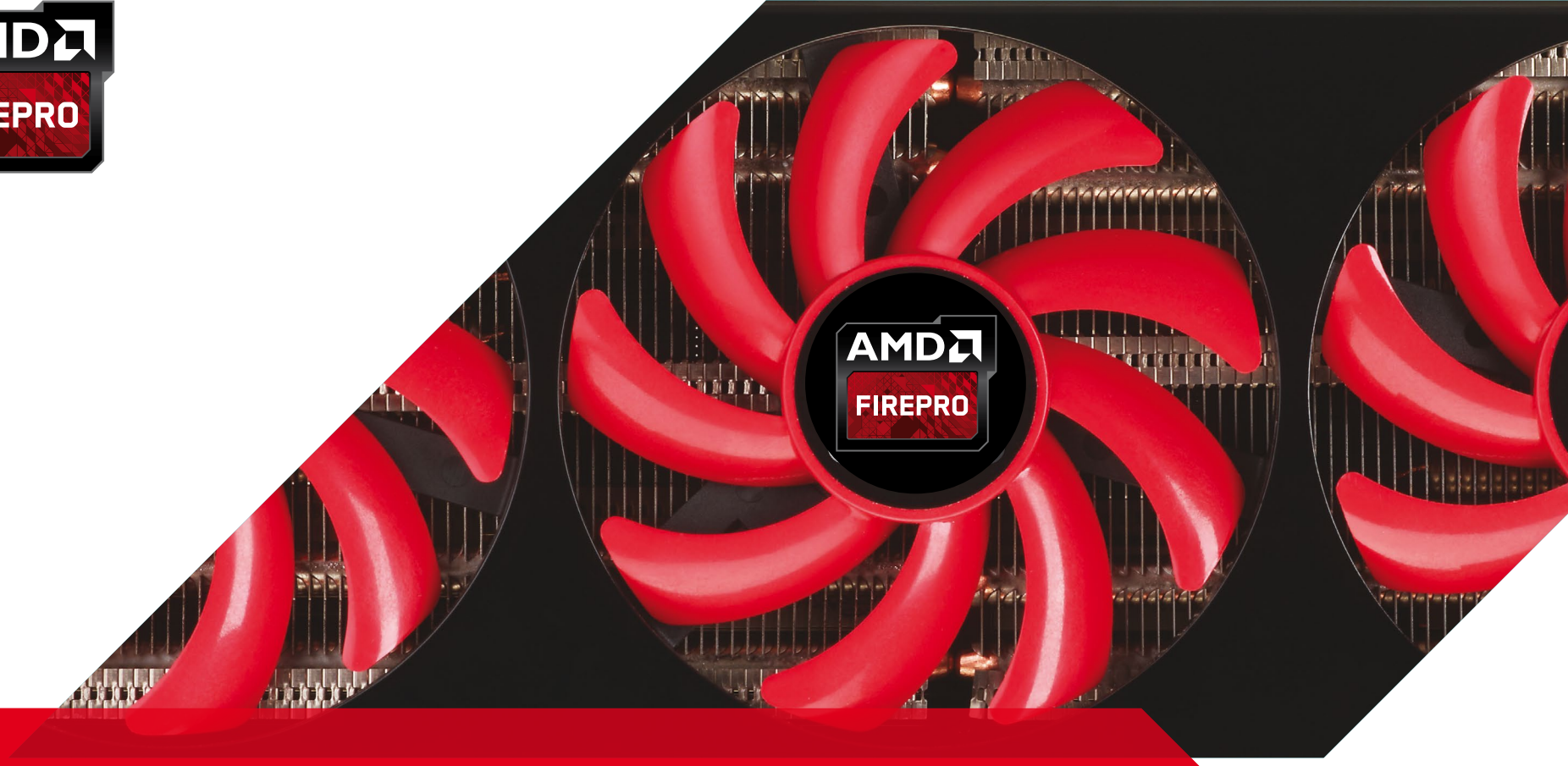
Several coding techniques have been discussed relating to the different implementations of our evaluation procedures. We thought it appropriate to clarify a few points about them.

• **Vectorization -** Changing an arithmetic operation on a scalar value into the same operation on a vector (fixed-size, preferably small) requires a complete refactoring of the algorithm and some additional branching management. The impact on code generation is heavy and performance of the generated code is highly dependent on the micro-library that makes use of intrinsics. Note that the effort has proved beneficial on both platforms. It should be worth your time on regular CPUs as well, especially when using AVX.

• **Alignment -** "Alignment" here stands for pointer dereferencing alignment. Memory should only be accessed by patterns of a certain size and at certain offsets. Enforcing alignment or "striding" is easy with small synthetic examples but often quite harder with real-life cases, especially when array indices derive from a complex computation. Our experiments showed that dynamically testing alignment with K20X comes at a reasonable expense. So does the use of gathers instead of aligned loads with Phi SE10P.

• **Read-only assertion -** When accessing memory, some assumptions can be made, for example that no other entity involved in the execution of the function/kernel will modify the data. In general, automating this assertion can easily be hinted by the programmer.

AMD FIREPRO



Be locked or be free

OpenCL

AMD FirePro™ S-Series solutions are the very best in open-standard GPU acceleration with OpenCL™ 1.2. Programmers can maximize their investment when developing code by targeting multi-core CPUs, the lastest APUs and discrete GPUs, freeing themselves from proprietary technologies.

Find out more @ www.amd.com

/hpc_labs

DISCOVERING OpenACC 2.0 – PART I

THE NEW DATA MANAGEMENT FEATURES

After a first release that proved essential to move parallel computations in C/C++ and Fortran to accelerators in a standardized way, OpenACC 2.0 was long due. Now that it's here, this series of three articles proposes to help you make the most of its new capabilities and expanded functionality. This month, we'll start with data management, an area that really needed improvement...

STEPHANE CHAUVEAU, PhD

The first OpenACC API specification (v1.0) was initially released in 2011 by PGI, Cray and NVIDIA with support from CAPS. At the time, it was introduced as a temporary test ground for the future accelerator extensions of OpenMP. Now, after almost two years of improvements by several new members, version 2.0 aims at becoming the *de facto* standard for directive-based programming of accelerators in C/C++ and Fortran.

One of the recurrent complaints about OpenACC 1.0 was the lack of flexibility in data management. Offloading data to the accelerator could be achieved by creating a so-called data region that had to be perfectly nested in the code. The **data** construct provides the most obvious way to create a data region but implicit ones also exist around compute constructs (**parallel** and **kernels**). The standalone **declare** directive can also be used to define

a data region corresponding to its current scope, which can be either local to the current procedure or global to the application. In any case, the rules to offload variables basically follow the scoping rules of variables in the source language.

This choice makes sense from the compiler standpoint since by construction the lifetime of a data region is entirely contained in the lifetime of the host variable it refers to. Such a strategy can however become problematic in applications using more complex data structures. OpenACC 1.0 does not provide any mechanisms to manage a dynamic number of data offloads except by using tricks such as creating a recursive function to encapsulate a data region as many times as needed. But that is clearly not an elegant solution. It is also impossible to create codes in which offloaded data has to be allocated and deallocated in different functions. This is typically the case in Object-Oriented applications where object constructors and destructors are the obvious places to manage the lifetime of offloaded data.

This last case is illustrated in **listing 1** where a function has to process by pairs an arbitrary number **n** of data structures provided by a vector of pointers. Since OpenACC 1.0 does not provide any mechanism to offload the **n** data structures simultaneously, the code keeps only two structures on the device at any time. This is of course quite inefficient since each of the **n** data structures has to be offloaded an average of **n/2** times. Performance could probably be improved by using the recursive trick mentioned above or by managing more than one structure per

Listing 1: Inefficient pair processing in OpenACC v1.0.

```
void doit( size_t n, struct Value * data[n] )
{
 for (size_t i=0;i<n;++i)
 {
  struct Value * data1 = data[i] ;
  #pragma acc data copy( data1[0:1] )
  {
   for (size_t j=i+1;j<n;++j)
   {
    struct Value * data2 = data[i] ;
    #pragma acc parallel copy( data2[0:1] )
    {
     work_on_a_pair( data1 , data2 ) ;
    }
   }
  }
 }
}
```

Listing 2: Dynamic offloading in OpenACC v2.0.

```
void doit( size_t n, struct Value * data[n] )
{
 for (size_t i=0;i<n;++i)
 {
  struct Value * data = data[i] ;
  #pragma acc enter data copyin(data[0:1])
 }

 for (size_t i=0;i<n;++i)
 {
  for (size_t j=i+1;j<n;++j)
  {
   struct Value * data1 = data[i] ;
   struct Value * data2 = data[j] ;
   #pragma acc parallel present( data1[0:1], data2[0:1] )
   {
    work_on_a_pair( data1 , data2 ) ;
   }
  }
 }

 for (size_t i=0;i<n;++i)
 {
  Value * data = data[i] ;
  #pragma acc enter data copyout(data[0:1])
 }
}
```

data region but that would require some aggressive changes to the original code, which is not the intended purpose of a directive based API.

Data management becomes dynamic

OpenACC 2.0 solves most of these problems by introducing new dynamic data management features. Two new standalone directives, **enter data** and **exit data**, can be used to respectively create and destroy offloaded data. Together, they are basically equivalent to a **data** construct but without the proper nesting requirements. Accordingly, they can be arbitrarily placed anywhere in the code and, unlike other of data constructs, they can even be executed asynchronously.

Listing 3: Custom implementation of a present_or_copy(A[n]) using API calls.

```
int A[n] ;

...

/* A[0:n] may or may not be present on the device */
int was_present ;
was_present = acc_is_present( A , n*sizeof(int) );
if ( !was_present ) {
  acc_copyin( A , n*sizeof(int) )
}

...

 /* A[0:n] is now present on the device */

 ...

 if ( !was_present ) {
   acc_delete( A , n*sizeof(int) )
 }

 /* A[0:n] may or may not be present on the device */

 ...
```

Listing 2 illustrates how the code in **listing 1** can be optimized in OpenACC 2.0. Each of the **n** data structures is now offloaded to the accelerator only once (one **copyin** and one **copyout**). The **enter data** and **exit data** directives are also provided as API calls (**acc_copyin**, **acc_copyout**, **acc_delete**, and so on).

If should be noted that since **enter data** and **exit data** do not provide a proper data region from the language standpoint, a **present** clause is still needed on the **parallel** construct to make the data accessible within the parallel region (this clause will not cause any kind of allocation or data transfer). The **enter data** and **exit data** directives and the associated API calls are especially therefore suitable for writing libraries in which data is managed and used by multiple functions.

Most of the data management features found in directives and constructs are now also provided as API calls. For instance, **acc_copyin()**, **acc_present_or_copyin()**, **acc_create()** and **acc_present_or_create()** correspond to the **data** clauses allowed on the **enter data** directive while **acc_copyout()** and **acc_delete()** correspond to the **data** clauses allowed on the **exit data** directive. These functions are unfortunately all synchronous but asynchronous alternatives will very likely make it to the next major OpenACC release, at least for a few of them.

Developers looking to implement their own data management features will also benefit from some low level API calls: **acc_malloc()** and **acc_free()** provide direct memory allocation on the current device. **acc_map_data()** and **acc_unmap_data()** associate arbitrary addresses on the host and on the device. **acc_deviceptr()** resolves the device address associated with a host address while **acc_hostptr()** performs the reverse resolution. Last, **acc_memcpy_to_device()** and **acc_memcpy_from_device()** transfer arbitrary memory regions between host and device.

Standalone data directives (and their equivalent API calls) can be mixed with data constructs but offloaded data created us-

ing one kind of directive cannot be destroyed using the other. Standalone data directives are not aware of each other so it would be incorrect, for instance, to re-implement the semantic of a **present_or_copyin** clause by just a call to **acc_present_or_copyin()** followed by a call to **acc_delete()**. The proper implementation, as illustrated in **listing 3**, requires the execution of both calls only if the data was not already present.

Implementing multidimensional arrays as vectors of pointers

It is common practice in C/C++ to implement multidimensional arrays as vectors of pointers. This is now directly supported by OpenACC 2.0 at least for regular bi-dimensional arrays. With a single data directive specifying a bi-dimensional shape, the OpenACC compiler will take care of the ugly details of creating an equivalent array of pointers on the device. An important restriction is that the pointers forming the array should all be present or non-present. Modifying the pointers either on the host or on the device is also not allowed.

This feature is directly applicable to our first example where the array of pointers **data[n]** can be treated as a bi-dimensional array **data[n][1]** of structures by OpenACC 2.0. This is illustrated in **listing 4** where the benefits in terms of simplicity and productivity are obvious compared to **listing 2**. Simple as it is, the syntax should not overshadow the fact that most

Listing 4: Multidimensional arrays offloading in OpenACC v2.0.

```
void doit( size_t nb, struct Value * data[nb] )
{
 #pragma acc parallel copy( data[0:nb][0:1] )
 {
  for (size_t i=0;i<nb;++i)
  {
   for (size_t j=i+1;j<nb;++j)
   {
    work_on_a_pair( data[j], data[i] ) ;
   }
  }
 }
}
```

Listing 5: Global variable with dynamic lifetime on the device.

```
//---- external.c ----

int counter ;

#pragma acc routine(update_counter)
void update_counter(void) {
 counter++ ;
}

//---- main.c ----

extern int counter ;

void work(void)
{
 #pragma acc parallel
 {
  /*
   * Problem: The compiler is not aware that
   * update_counter is accessing counter and
   * since counter has a dynamic lifetime on
   * the device, how can update_counter figure
   * out its (non-constant) device address?
   */
  update_counter() ;
 }
}

int main(void)
{
 #pragma data copy(counter) ;
 {
  work() ;
 }
}
```

implementations will actually generate one data transfer per pointer. Consequently, the performance of a non-contiguous array is likely to be significantly lower than that of a contiguous array. An interesting side effect of managing the array as a whole could however impact efficiency in a positive way: the loops used to iterate over the pointers can now be executed on the device, which may incidentally provide more hardware parallelism.

Support for arbitrary complex data types such as chained lists, trees or any other structure containing pointers is scheduled for futures OpenACC versions. In the meantime, managing complex dating is already possible using the v2.0 API calls.

Listing 6: Sample I/O library using the new link feature.

```
//---- io.c ----

#pragma acc declare copyin(io_status, io_size)
#pragma acc declare link(io_buffer)
int  io_size   = 0 , io_status = FALSE ;
char io_buffer[10000000] ;

#pragma routine(io_putchar)
void io_putchar(char c)
{
  if ( io_status == TRUE ) {
    io_buffer[ io_size++ ] = c ;
  }
}

void io_start()
{
  io_status = TRUE ;
  io_size   = 0 ;
  #pragma acc update device(io_size, io_status)
  #pragma acc enter  data create(io_buffer)
}

void io_stop()
{
  io_status = FALSE ;
  #pragma acc update self(io_size) device(io_status)
  #pragma acc exit data copyout(io_buffer)
  fwrite(io_buffer, io_size, 1 , stdout) ;
}


//---- main.c ----

void main()
{
  io_start() ;
  #pragma acc parallel
  {
    ...
    io_putchar('X') ;
    io_putchar('Y') ;
    io_putchar('Z') ;
    ...
  }
  io_stop() ;
}
```

Global data

Last but not least, support for global data was improved in the **declare** directive. In OpenACC 1.0, it was already possible to create a data region with a global lifetime using the regular **data** clause (*e.g.* **create**, **copy**, **copyin**...), while the **device_resident** clause was available to create device variables without a version on the host. Global variables with a dynamic lifetime on the device were supposed to be handled like all other variables by the implicit or explicit data regions enclosing the compute constructs. OpenACC 2.0 is now officially supporting calls to user-defined functions in the compute constructs and this is a problem because the compiler, when processing these compute constructs, may not be able to figure out which global variables with a dynamic lifetime are used by the called functions. The context is illustrated in **Listing 5** where a non local function called from within the compute construct is using a global variable with a dynamic lifetime on the device.

OpenACC 2.0 provides a simple solution to the problem by introducing a new **link** clause on the **declare** directive. A global variable declared as a **link** variable will always be present on the device in the form of a global link initially pointing to nothing – think of it as a pointer initially set to **NULL**. The link will be updated dynamically to point to the real address or to **NULL** every time the global link variable is managed by a **data** construct or by an **enter data** or **exit data** directive. An important restriction about global link variables is that they cannot be managed using API calls such as **acc_create()** or **acc_copy()**. The reason is that an API call is only given a host address and therefore cannot be aware of the link nature of the corresponding variable. So, logically, it cannot update the associated link on the device.

One major drawback of global link variables over regular variables is that it is possible but of course illegal to access them on the device while they are currently not managed by any data region. Such errors are usually impossible to catch at compile time and in most cases may cause a runtime error similar to that of using an illegal pointer on the device. Anyway, to illustrate the use of the **link** clause, **Listing 6** shows an oversimplified I/O library in which the huge output buffer is dynamically managed as a **link** object.

You should now be able to take advantage of OpenACC 2.0's new data management features - and implement them in your codes - easily. Next month, we'll focus on interesting improvement within the compute constructs: atomic constructs, device specific clauses, nested parallelism, routine calls, loop tiling, etc. Stay tuned and...

Happy programming! ■

Subscribe now!

[for free, forever]

Subscribe now and receive, every month,
an expert, actionable coverage of HPC and
Big Data news, technologies, uses and research...

+ get yourself access to exclusive contents and services

www.hpcmagazine.com