MARCH 2014

HighPerformanceComputing

The reference media in HPC and Big Data

Volume I | Number 3

Americas

0x03

/success-story

DEM MODELING

Complex particles domesticated

/hpc_labs

DISCOVERING OpenSPL

Is spatial programming the future of HPC ?

/state-of_the_art

SHIELDS TO MAXIMUM!

Simulating the impacts of orbital debris

/hpc_labs

PERFORMANCE PROGRAMMING

A scientific, practical methodology for measurable results (part II)

THE UBERCLOUD HPC EXPERIMENT: AN INSIDE LOOK

How - and why - HPCaaS is becoming a reality

An HPC Media publication

Volume I, number 3

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

From the Publisher

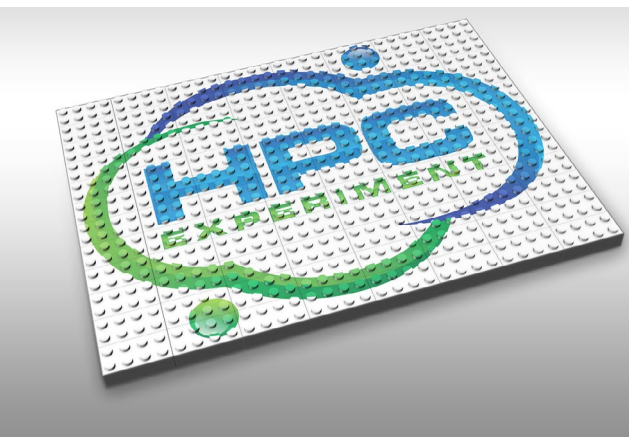
Dear readers,

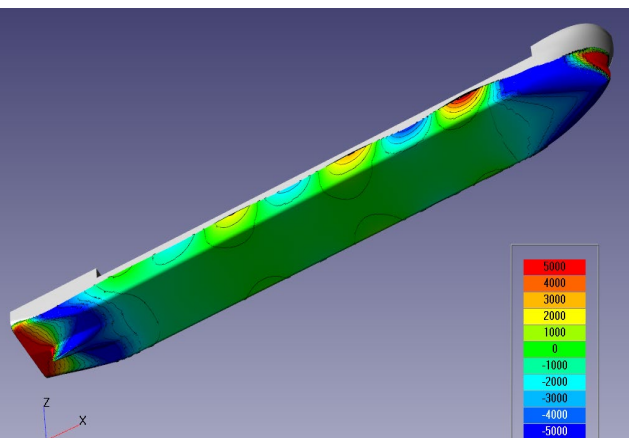
when do you plan to run your computation / simulation jobs in the Cloud? The question, pushy as it may sound, has never been more timely. For this month's cover story, prepared in collaboration with the co-founders of the UberCloud HPC Experiment themselves, leaves no room to doubt. Given the level of maturity reached by the technological building blocks, given the virtually endless possibilities offered by infrastructure elasticity, given the indisputable financial benefits of on demand resources for most non-academic sites, it is clear that the future of supercomputing is largely in the Cloud. You'll see...

Besides remote HPC and its present pros and cons, this third issue of HighPerformanceComputing covers a wide variety of topics, from an off-road example of discrete elements modeling to the simulation of orbital debris impacts on space vehicles. Space, again, is the raison d'être of OpenSPL, a new programming paradigm using spatial locality to raise performance and energy efficiency to new levels. It is presented to you by its conceptors. Last, those of our readers who appreciated our in-depth introduction to performance programming will be happy to find it completed this month with a section entirely dedicated to the parallelization of existing codes.

Happy reading!

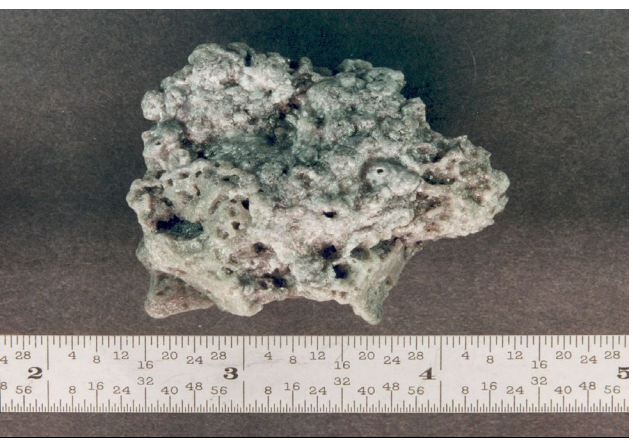
frederic@hpcmagazine.com

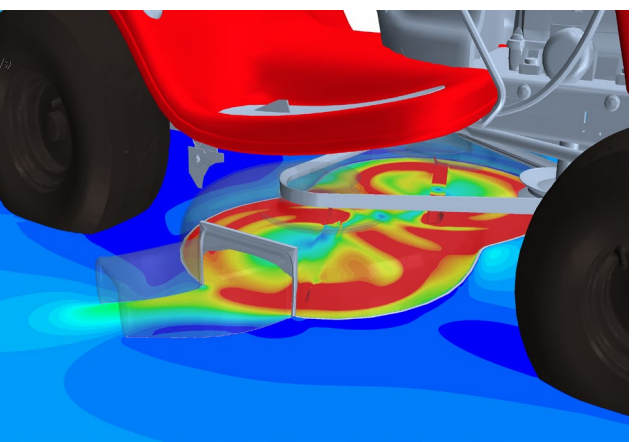




CONTENTS

THIS MONTH





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

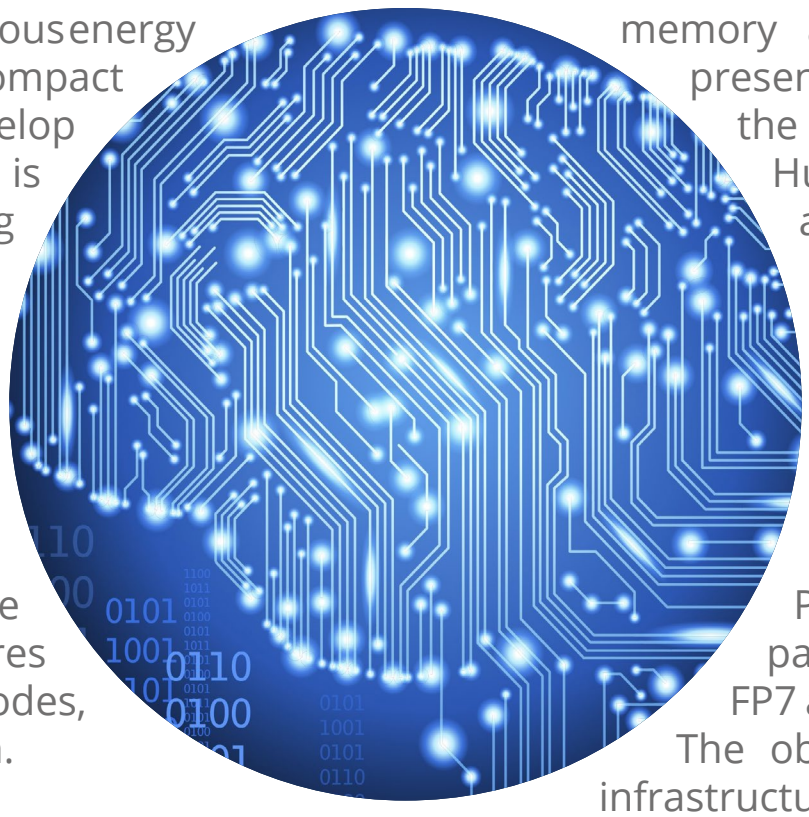
CRAY®

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

/NEWS

Human brain *vs* von Neumann at ISC14

As Professor Karlheinz Meier, of Heidelberg University, emphasizes, the brain is characterized by its tremendous energy efficiency, fault tolerance, compact size, and ability to develop and learn. Ultimately, this is what current computing systems are required to have, an expectation that will only grow as exascale systems come closer to reality. This parallel may appear utopian, but in light of advances in the neurosciences and the particular technical features of new VLSI computing nodes, this reality is on the horizon.



In any event, this is the idea Prof. Meier intends to promote in an ISC keynote this coming June. What's more, he will highlight the differences between such a system and von Neumann's which is based on a conventional processor-memory architecture. He will also present international projects in the field, namely the European Human Brain Project, which is attempting to build a direct link between biological data, computer simulations, and configurable hardware implementations derived from neuromorphology. Spanning a ten-year period, the Human Brain Project is being financed as part of the European Union's FP7 and the Horizon 2020 plans. The objective is to build a new infrastructure for neurosciences and brain research in a medical and information technology context. Rest assured we will be there and, by proxy, you as well.

Bristol labeled Intel Parallel Computing Center (IPCC)

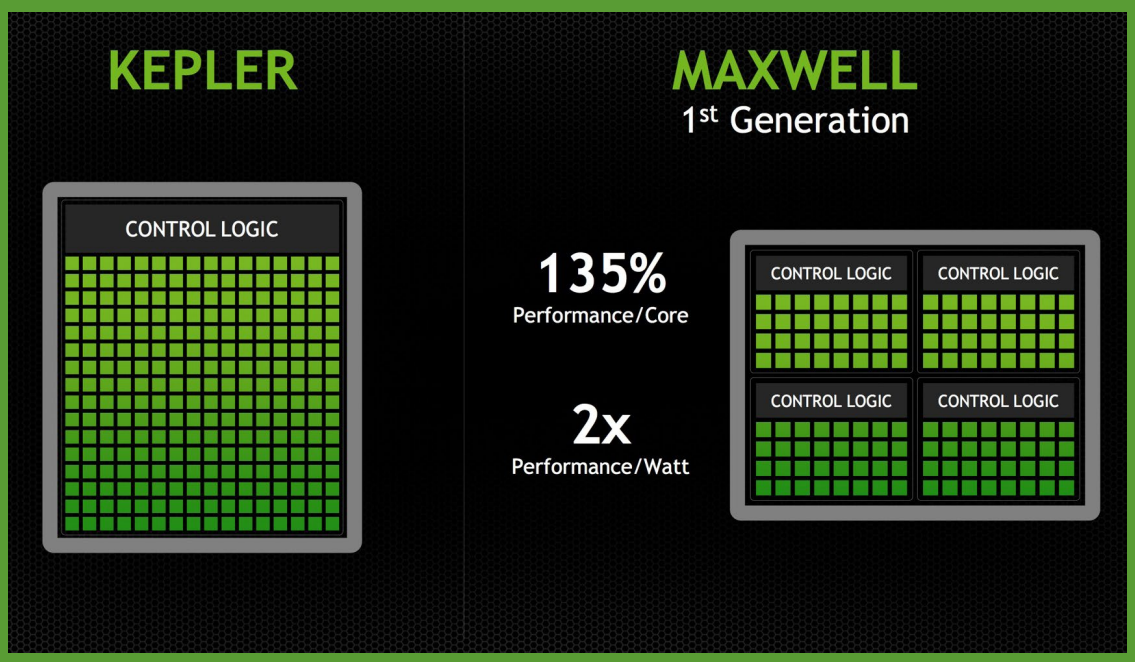
After the TACC, the Universities of Tennessee and Purdue, ICHEC, ZIB, and CINECA, Bristol University in Great Britain has become the seventh Intel center dedicated to parallel computing, and is now labeled an IPCC. These centers' vocation is to modernize current codes to make them more parallel-capable, so that modern processor architectures can be put to better use. They also aim to train students, scientists, and researchers in parallel programming techniques, which is a whole other challenge in itself.

Intel chose Bristol University for its research in manycore parallel architecture usage and for its leadership in standard and open models in parallel programming. This falls into the category of exploratory academic work that the chip manufacturer finances generously. Moreover, IPCCs are part of a program in which the more concrete objective is to address current challenges in parallel computing in the fields of energy, industrial science, life sciences, etc. As such, they receive funds for a two-year period which can be renewed, depending on provable results. Bristol, in particular, will be in charge of optimizing applications in the fields of fluid dynamics and molecular engineering.

NVIDIA Maxwell: in total discretion...

We're not accustomed to a quiet release of a new GPU architecture from NVIDIA. Yet this is what has happened with the launch of the GeForce GTX 750 Ti card, which introduces the Maxwell architecture, successor to Kepler and eagerly awaited by the HPC community but perhaps less so by gamers. And for good reason, as this new discrete GPU will not surpass records in terms of image resolution.

It will, however, double Kepler's energy efficiency - a significant advance for the mobile market (notebooks and tablets). A practical advantage of this efficiency is, for example, the absence of an external PCI-Express connector to power the GPU: the energy comes from the card to which it is connected which, incidentally, will simplify the design of hosting hardware .



Now, some of Maxwell's features are worth mentioning. Overall performance is increased by 35% for a 1:4 DP/SP ratio. The main innovation comes from the multiprocessors' architecture. New SMMs (streaming multiprocessors) replace Kepler's SMXes and are now cut into four smaller blocks supporting 32 CUDA cores each, or 128 on an SMM as opposed to 192 on an SMX.

This granularity combined with better workload distribution improves the efficiency of parallelism. The occupancy is thereby doubled, provided that the use of registers and shared memory is not a limiting factor. Another notable feature is the larger L2 Cache (up to 2MB). Last, unified memory support is now at the hardware level - another important improvement over Kepler.

Dassault Systèmes + Accelrys, a chemical marriage

By launching a friendly take-over bid on Accelrys, a scientific software solution publisher based in San Diego, Dassault Systems' aim is to enrich its product lifecycle management solutions in the fields of molecular chemistry, from discovery to the production phase, in the life sciences industries, consumer products, high technology, and energy. The company is also acquiring a prestigious client portfolio consisting of the likes of Sanofi, DuPont, Unilever, and L'Oréal.

This majority stake will indeed allow Dassault to broaden the spectrum of scientific and technological fields targeted for the development of new products, from original idea to final design. Accelrys already provides PLM solutions with a software portfolio allowing scientists to access, organize, analyze, and share data in order to improve their productivity. The bid, approved by the Accelerys Board of Directors, is for all outstanding shares converted into cash for a valuation of $750 million.

Lenovo's new deal

Lenovo's acquisition of IBM's x86 server division is being seen as a major disruption in the OEM server market. And indeed, based on the current Top500 figures, it's easy to realize the impact. In last November's list, the number of IBM machines accounted for about one third of the total number of systems, the same as HP.

In terms of combined performance, this represents 32% of the total power, *vs* 16% for HP. Now, let's take a look at x86 computers. They account for approximately 77% of the machines on the list, but only 34% of the cumulative power. BlueGenes and other amply-sized Power-based systems are indeed the ones with the most powerful configurations. The result is that by buying IBM's x86 server business, Lenovo is acquiring a choice position at the top of the chart. All other things being equal, Lenovo would come in second place behind HP, with 25% of the listed systems, but would drop to fifth place in terms of overall performance.

How things will play out in the long term remains to be seen. Although Lenovo's know-how is unquestionable, the fact that it is not American could cause it to lose some public and private contracts in the US.

PGI compilers: the 2014 collection has arrived!

The latest versions of the PGI development tools (which have been under the NVIDIA banner since last year) are revealing some interesting new features. In addition to the anticipated support for the Tesla K40 accelerator with version 2.0 of the OpenACC instruction set, PGI is also opening OpenACC up to AMD's APUs and Radeon discrete GPUs. Those who were worried about the publisher's independence after the buyout can lay their fears to rest! Another noteworthy advancement is the interoperability of the CUDA Fortran and OpenACC version with the Allinea DDT debugger. Lastly, our readers who develop on the Mac will be pleased to find a now free version of the Fortran 2003 and C99 compilers for that platform.

This 2014 crop is thus improving the portability of code to various acceleration platforms, a specialty that PGI shares with the CAPS OpenACC compiler which, for its part, also supports Intel's Xeon Phi. It also allows everyone to benefit from OpenACC 2.0's important advancements, namely global and dynamic data management. For more information on this last topic, essential to code acceleration, feel free to refer to two in-depth articles which your favorite magazine has just devoted to the subject.

China will have its Power8

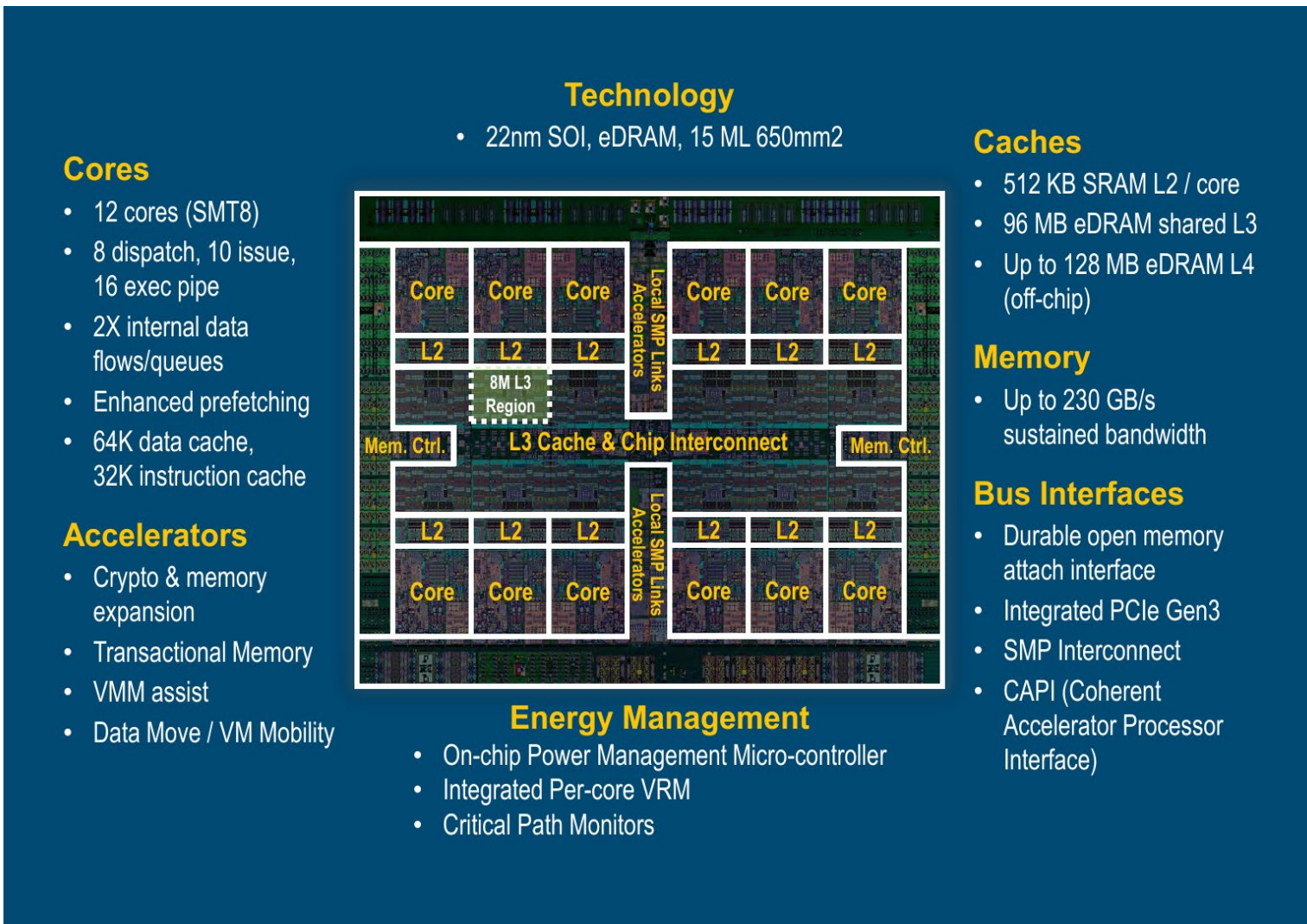
The mutual interest was obvious. On one end, IBM wanted to extend the realm of the OpenPower Consortium as far and wide as possible, because the future of high performance computing does not necessarily lie in x86. On the other end, the Chinese government was seeking to occupy every space where the future of high performance computing is playing out. Which is why China will have its Power8 processor, with IBM granting an industrial license for its architecture to a company created for the occasion, Suzhou PowerCore.

Similarly to what ARM does (intelligently) to unite a community of hardware and software developers, to develop an ecosystem and to ensure that its architectures live on, IBM continues to open up to businesses interested in the new iteration of Power. Until now, there haven't been very many major participants. In addition to Google, which is always in the right place at the right time, NVIDIA and Mellanox are also in the mix. But the arrival of a Chinese designer adds another dimension, especially since the agreement specifies that Suzhou PowerCore may modify the Power8 plans as it sees fit and have the resulting processors manufactured where it pleases.

As our regular readers know, the Beijing Academy of Sciences, which serves as a governing body for HPC in the Celestial Empire, is currently working on no less than six different processor architectures. Which is further evidence of the importance the Chinese attribute to high performance computing. With Power8, the objective, however, is more commercial than technical or strategic. According to Suzhou PowerCore officials, the projects targeted by these processors, aiming for the first half of 2015, are less oriented toward research than "large-scale analytics" and Big Data.

The first versions of Power8 - all developers combined - are scheduled for September of this year. IBM spokespeople have announced 12 cores running at 4 GHz with 8 threads per core, 96 MB of L3 cache on the die and 128 MB of L4 cache in the memory controllers, for a sustained memory bandwidth of 230 GB/s and an I/O bandwidth of 48 GB/s, that is, approximately twice as much as the Power7+.

Since Suzhou PowerCore is a subsidiary of C*Core, another Chinese company which has produced more than 90 million SoCs (ARM, etc.) so far, the Power8 made in China should be widely distributed. What benefits it will have over its Western competitors now remains to be seen.



/news

HPC in 2014 : the major trends according to IDC

As is customary every year at this time, IDC has just published its ten major predictions for the HPC market in 2014. Those familiar with IDC know that this document is more like an in-depth analysis of our ecosystem and its trends than a crystal ball game.

And it just so happens that these trends are occurring in an evolving landscape. It starts with the now nearly universal recognition of high performance computing as an economic and strategic value giving the industry a clearly positive outlook for the longer term. Additionally, although Big Data is booming ($16.1 billion forecasted for 2014), its relevance for most businesses has yet to be established. Also worthy of note is the confirmed takeoff of Cloud Computing, as witnessed by the constant increase in offerings and players with a broad range of service types. Lastly, Lenovo's very recent acquisition of IBM's x86 server business is being seen as a major disruption of the dynamics in the OEM server market.

Now let's get into details. The High Performance Computing server business declined by nearly $1 billion in 2013 after three record years, but should bounce back in 2014 and return to levels just short of the 2012 figures, in volume or number of units shipped. IDC estimates that Lenovo's arrival on the scene should lead to some degree of market share redistribution. So far, IBM and HP each account for about one third of this market, with Dell in third position at about 15%. According to IDC, Lenovo should be capable of matching Dell's share.

As for supercomputers, IDC expects to see 100-PFlops systems between late 2014 and early 2015 in China, the US, Europe (PRACE), and Japan. And when it comes to exascale systems, analysts think that their roll-out in 2020 will not be impacted by any disruptive technology, and that it is now just a matter of investment. Similarly, the study is projecting a power consumption of 20 to 30 MW, but does not expect this energy efficiency to be reached before 2022-2024.

While a high-end supercomputer currently costs between $200 and $500 million - remember that price tags rarely exceeded $100 million just 3 to 4 years ago - IDC confirms that the amount could reach one billion in the next three years. Understandably, at these levels, ROIs will have to be justified! We can therefore expect an explosion of communication on achieved results, be they scientific, economic or even societal. In addition to ROI, technology transfers and competitiveness are among the main reasons why partnerships between computing centers and industry, such as PRACE and INCITE, are on the rise.

On the electronics end, IDC notes that, although coprocessors and accelerators have become mainstream (77% of high performance computing sites are so equipped versus 28% in 2011), they are still widely used for experimental purposes. So the dominance of x86 processors is not being threatened, not even close, as they still account for 80% of sales.

For IDC, the main reason for this roadblock to accelerator adoption is programming difficulty, or at least the necessary additional effort. The issue of the time it takes to program or reprogram an application in anticipation of the expected performance improvement is a deterrent to migration. But, somewhat surprisingly and without providing any decisive arguments, IDC expects major leadership changes in the CPU market.

IDC has not forgotten storage and interconnects, which are essential elements in a data-saturated world. Sales derived from storage, the most dynamic segment of the HPC market, should reach record highs ($6 billion by 2017). However, with the flow of information being a difficult yet crucial issue, the interconnects market is in full transition, a transition gradually leading it away from the traditional computation-centric model.

And lastly, finishing with the Cloud, the number of sites using on-demand services for their HPC needs has grown from 13.8% in 2011 to 23.5% in 2013, making it a genuine trend. Here, too, the roadblocks to adoption are known: data security, transfer time, performance of highly parallel loads and application-defined infrastructure. But the major players in this industry have now taken these roadblocks into account, resulting in a gradual emergence of offerings and solutions intended to solve these problems. So you can bet the real figures for 2014 will show the same progression dynamic.

IDC Top 10 HPC Predictions for 2014

1. HPC server market growth will continue in 2014, after a decline in 2013
2. The global exascale race will pass the 100PF milestone
3. High performance data analysis will enlarge its footprint in HPC
4. ROI arguments will become increasingly important for funding systems
5. Industrial partnerships will proliferate, with mixed success
6. x86 base processor dominance will grow and competition will heat up
7. Storage and interconnects will benefit as HPC architectures gradually course-correct from today's extreme compute centrism
8. More attention will be paid to the software stack
9. Cloud computing will experience steady growth
10. HPC will be used more for managing IT mega-infrastructures

IDC
Analyze the Future

"The value of an idea lies in the using of it"

The famous quote by Thomas Edison applies perfectly to the new computer at the NERSC (Berkeley) bearing his name. Although the highly energy-efficient machine has a "modest" theoretical peak computing capacity of 2.4 Pflops, it was designed not only to perform large-scale simulations and modeling, but also and predominantly to process the enormous datastream they generate. *"DoE sites are inundated with data that researchers cannot adequately process or analyze,"* explains Sudip Dosanijh, Director of the NESRC Division. *"That's why Edison was optimized for these two types of computing, which require rapid management of data movements."*

Indeed, Edison, a Cray XC30, was developed with ultra fast interconnectivity, a wide memory bandwidth, a lot of memory per node, and very quick access to the file system and disk drives. In fact, it is the successor to a Cray XE6, which had become somewhat obsolete given the specificities of these tasks. As it does not include accelerators, no code needs to be rewritten, which makes the machine immediately available for simulations, a time saver clearly appreciated by the lucky scientists who will work with it.

In addition to this computational capability and immediate availability, Edison incorporates an advanced "free cooling" (ambient air cooling) technology also developed by Cray. The principle of this system, free of mechanical workings, is based on air flow between cabinets rather than through the cabinet from the door to the back. Concretely, each cabinet contains a heat sink that cools the air before it moves on to the next cabinet. According to Cray, the machine achieves a PUE of approximately 1.1, a figure that is indeed easier to reach with a passive approach.

In figures, Edison features 2.57 Pflops of peak power, 357 TB of aggregate memory, 5,576 nodes dispatched in 30 cabinets, 133,824 Intel Ivy-Bridge processor cores (12 cores per compute CPU), 7.56 PB of disk storage, 462 TB/s of bandwidth for global memory, 163 GB/s of I/O bandwidth, and 11 TB/s of bidirectional network bandwidth.

/news

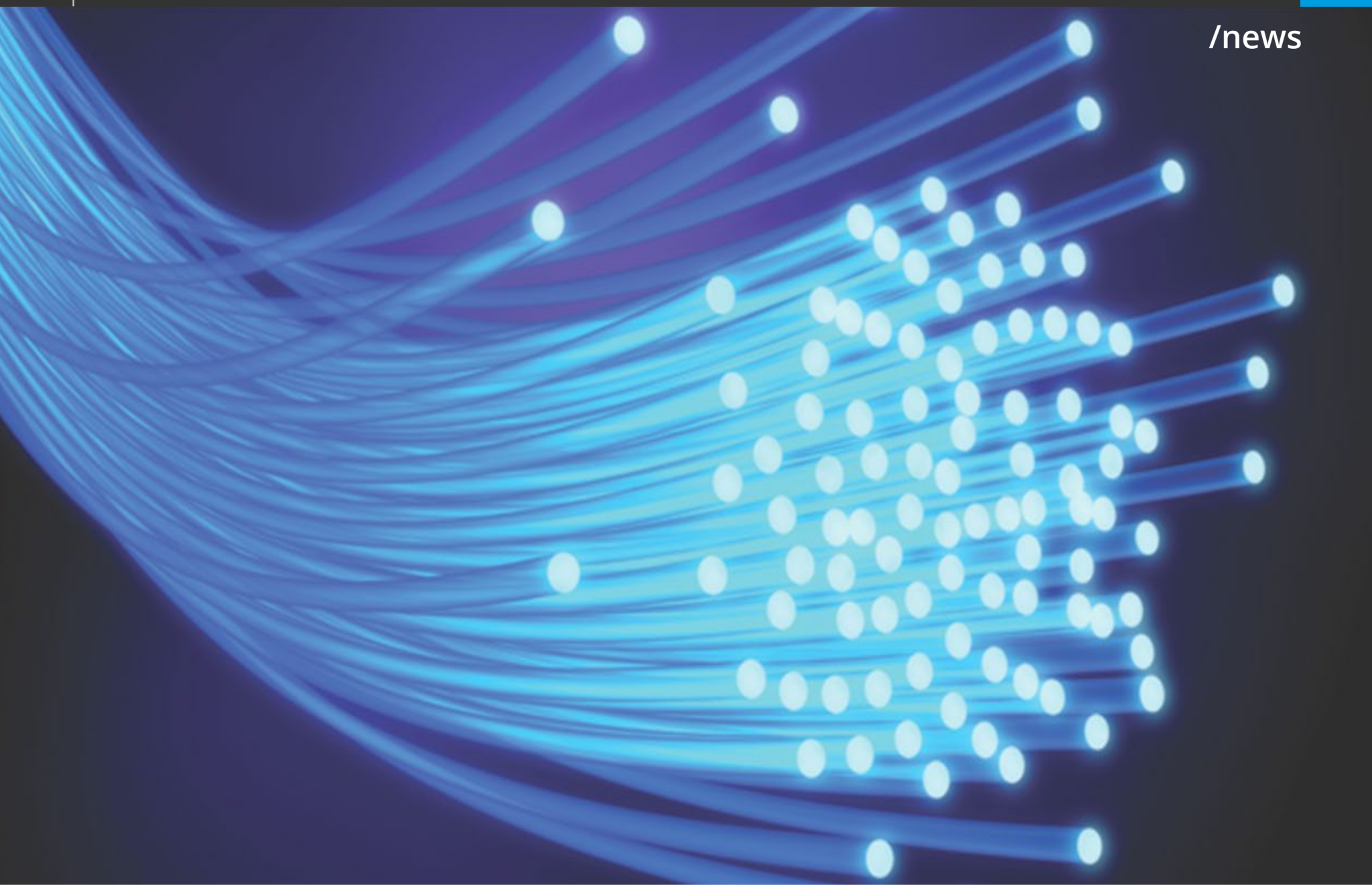
Xeon E7 v2 : the target is the Big Data!

After three long years, this refresh of the Xeon E7 8800/4800/2800 family was highly anticipated. Based on the Ivy Bridge EX architecture, which is now manufactured with a 22 nm process, this new generation is further distinguished by 50% more cores.

The main point however is that it was primarily designed to meet a specific application objective, i.e. the real-time processing of massive quantities of data. To that end, the memory capacity is tripled compared to the previous generation, with up to 1.5 TB of DDR3 per socket: ideal for storing large databases and eliminating costly disk access! Obviously, 6 or 12 TB in machines having 4 or 8 sockets (32 is the maximum number of sockets per machine) will allow data-intensive applications to benefit from better latencies in NUMA environment than if they had to run on a cluster using Ethernet or InfiniBand, given the same capacity in terms of the number of memory sockets.

With 50% more cores (for a total of 4.3 billion transistors per die), each processor can contain up to 15 cores, compared to 10 with the previous generation. Why 15 instead of 16, a logical multiple of two for binary systems? Rumor has it that a backup core is reserved for Intel's new Run Sure technology.

As for configurations, 2, 4, and 8 processors can be QPI-linked in the same system. As we go to (digital) press, some OEMs, such as HP or SGI, are designing their own controllers (and chipsets) to accommodate more processors, and a certain number of other big guys may follow. Now, regarding the thermal envelope, it appears to be remarkably well controlled as it should not exceed 155 watts in the hottest configurations. The pricing? In the neighborhood of $6,500 per unit. For the record, Intel is claiming 80% better performance than IBM's POWER7+ for an 80% lower TCO over four years, thereby reducing the RISCs, so to speak...



1,4 Tbps sustained over a commercial network!

Month after month, records get broken. At the same time, it's precisely what they're made for... At ISC 2013, we reported on Alcatel Lucent's and T-Mobile's successful experiment, which consisted in sending a sustained 400 Gbps dataflow between two climatology and industrial digital simulation applications. This was already far beyond the standards at the "time," even though the communication had been conducted under near-experimental conditions. Well, we just learned this month that Alcatel has done it again, in partnership with British Telecom, this time with a sustained flow rate of 1.4 Tbps - in itself nothing to sneer at - on an existing production network. You read correctly: neither the hardware nor the fibers had to be changed.

The achievement took place in October 2013 on a fiber-optic network connecting the BT Tower in London to the Adastral Park University campus in Suffolk, but it was not made public until this month. To disbelievers who would object that this performance is not such a big deal because of the pull of gravity along the North-South direction, let us add that the flow rate was achieved in both directions! Seriously, Alcatel said that all it had to do was reprogram its hardware platforms. It is therefore an important advancement from an industrial standpoint. For example, Huawei had demonstrated a 2 Tbps rate over the Vodafone network a few months ago, but with substantial modifications to the existing infrastructure.

Recordswise, the bar is now set to 10 Tb. In the laboratory, Alcatel successfully sent 31 Tbps over an underwater fiber measuring 7,200 km long, with optical amplifiers every 100 km, while NEC and Verizon have reached 40.5 Tbps over 1,800 km of "conventional" fiber. Each of these two innovative companies is hinting at new announcements for SC 2014 which, let's not forget, is also the high performance communications conference...

Bull refocuses on its added value

While, for financial analysts, the objective of the new "One Bull" plan is to double the group's operational performance to 7%, for the rest of the world - namely associates and customers - the ambition of the announced strategy is to refocus Bull on its business with the highest added value. The communication people put forward the Cloud and the Big Data, two catch terms in a context where Bull is clearly positioned to become the "trusted operator for business data." But the details of the plan reveal a more structure-centric reorganization hinging around the two dimensions of "data infrastructure" and "data management," with all that implies for technical team redeployment.

Indeed, One Bull, to be implemented until 2017, is based on three pillars, which can be considered logical in light of the group's expertise: high performance computing and large infrastructures, complex software integration, and data security.

As such, Bull wants to simplify its internal organization with a substantial reduction in its number of operational units, a global geographic grouping around five regional hubs (instead of the current 50 subsidiaries), and a unified sales organization. It must be stressed that this strategy does not include a layoff plan and that, at the same time, it earmarks a large proportion of revenue - 6% - to technical and growth investment via the "iBull" program. Not common currency these days.

Mellanox publishes its Ethernet Switch API

It's done. Mellanox just released its Ethernet Switch API. The Software Development Kit (SDK), with full support for Layer 2 and Layer 3 functionality, can now be downloaded from GitHub. For Mellanox, this move represents a contribution to the Open Ethernet initiative and an additional commitment to the Open Compute Project. The release is intended to enable the community to develop Open Source network applications, optimize application communication protocols and establish a solid foundation for the development of a standardized switching interface.

Now, let's put things in context. Customer needs - public or private - regarding networking are changing, especially with the emergence of SDN (software-defined networking) and the rise in power of OpenFlow. The fact that Mellanox, uniquely positioned in distributed and I/O-intensive high performance communications, is actively involved in the establishment of an open standard for Ethernet can only strengthen its legitimacy as a provider of more demanding solutions. Down the road, the winner may very well spell "InfiniBand".

Microsoft opens its cloud

It is quite exceptional to be noticed: Microsoft participates in an Open Source event. No, Windows' source code is not being released (yet) but, by becoming a member of the Open Compute Project created by Facebook three years ago, Microsoft is opening the infrastructure design of its Windows Azure cloud.

Although a major player of the likes of Amazon and others, Microsoft probably does not want to miss the boat on this promising sector: "*We do this because we want to lead innovation in the design of the Cloud and of datacenters*" said Bill Laing, Corporate VP of Cloud and Enterprise during his keynote at the Open Compute Summit in San Francisco. "*And of course, we also want to learn from the community...*"

For the record, the design of the servers that host the Windows Azure public cloud shows significant economies of scale. The chassis, for instance, would help save 10,000 tons of metal and 1,700 km of cables, simply by being modular. Without going into detail, each rack can be configured as a compute or storage node. They can plugged or removed without wiring which, for centers hosting hundreds of thousands of servers, represents a substantial time saver.

ARCHER up and running!

Like NESRC, which recently celebrated the launch of the Edison machine, the University of Edinburgh welcomes its own CRAY XC30, christened ARCHER, due to replace HECToR (a CRAY XE6, like Edison's predecessor, that was put in production just seven years ago).

ARCHER, short for Academic Research Computing High End Resource, is part of the resources available to the academic research of the British Crown and connected to the UK Research Data Facility network. The machine, totaling 1.56 Pflops peak, features the latest 12-core Intel Xeon E5 v2, two in each of the 3008 nodes dispatched in 16 cabinets. The CPUs are connected through the CRAY Aries network interface. As at NESRC, scientists can already use the machine, simply by recompiling their applications with the provided suite of tools, namely compilers from CRAY, Intel and GNU.

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

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

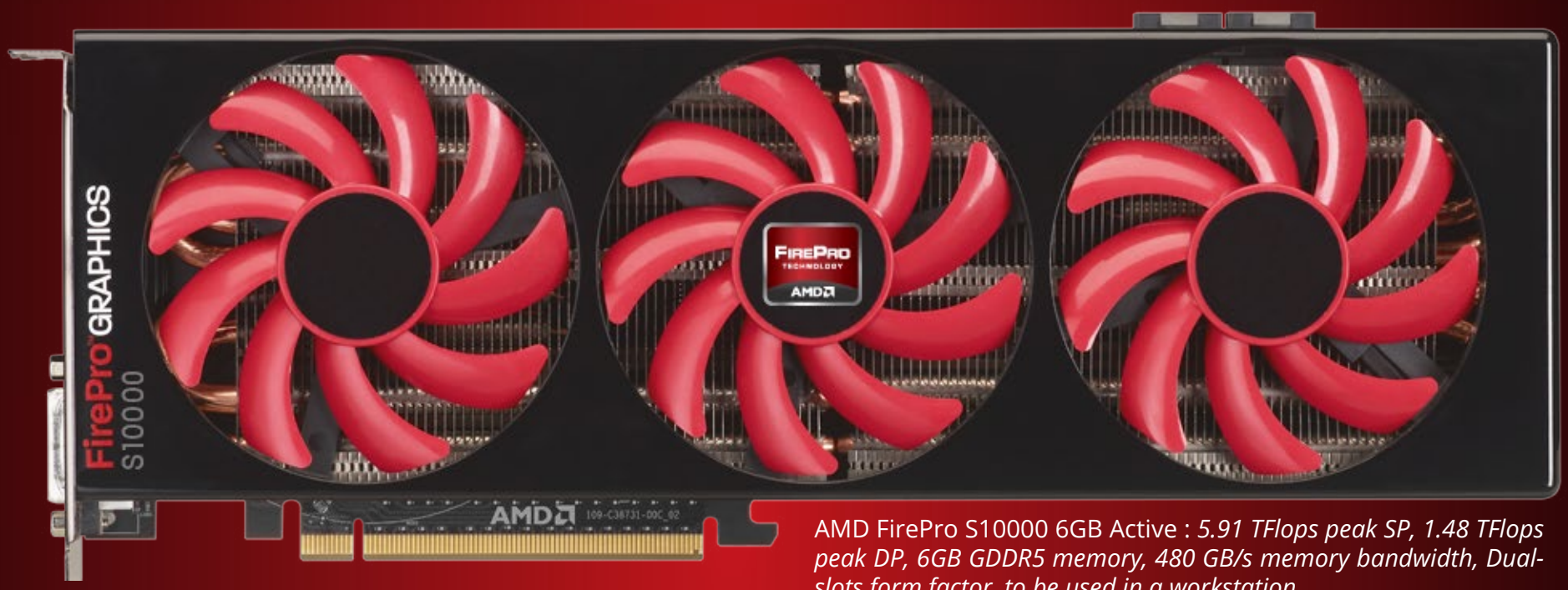
Fits comfortably in OpenCL developers workstations.

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.



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

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.

Last month's winners:

Marcin Ziolkowski, Clemson University
Gregg Charles, GX Technology

The operation is now closed.

/AGENDA

MARCH 2014

GPU Technology Conference
Where: San Jose, CA, USA
When: March 24-27
GTC is the world's biggest and most important GPU developer conference. It offers unmatched opportunities to learn how to harness the latest GPU technology, along with face-to-face interaction with industry luminaries and NVIDIA experts. Stay tuned for announcements on who'll be bringing the 'wow factor'...

28th Annual HPCC-USA Conference
Where: Newport, RI, USA
When: March 25-27
Supercomputing: What Does the Future Hold? Welcome to the next conference of the National High Performance Computing and Communications Council...

APRIL 2014

EASC2014 - 2nd Exascale Applications and Software Conference
Where: Stockholm, Sweden
When: April 2-4
This conference brings together developers and researchers involved in solving the software challenges of the exascale era. The conference focuses on applications for exascale and the associated tools, software programming models and libraries.

4th Cloud & Big Data Summit
Where: London, UK
When: April 10-11
The 4th Cloud And Big Data Summit 2014 will focus on the emerging area of cloud computing enhanced by latest developments related to infrastructure, operations, security, governance and services available through the global network. Panels, sessions and workshops are designed to cover a range of latest topics, trends and innovations related to cloud computing and data management. This Conference is a European platform to identify your company as a key player in the dynamic game changing market priorities in the field of cloud computing, data and information security.

MAY 2014

2nd ASE International Conference on Big Data Science and Computing
Where: San Jose, CA, USA
When: May 27-31, 2014
The Second IEEE/ASE International Conference on Big Data Science and Computing aims to bring together academic scientists, researchers, scholars and industry partners to exchange and share their experiences and research results in Advancing Big Data Science & Engineering.

JUNE 2014

6th Annual Cloud World Forum
Where: London, UK
When: June 14-18, 2014
Cloud World Forum is EMEA's leading Cloud event with the industry's most comprehensive agenda of all things Cloud.

ISC'14
Where: Leipzig, Germany
When: June 22-26, 2014
Join the global supercomputing community, ISC'14 again brings together 2,500 system managers, researchers from academia and industry as well as developers, computational scientists and industry affiliates from over 50 countries.

To our readers

In last month's BloodHound SSC Success-story, we failed to mention that the interview was originally published by **Renuda**. We apologize for the oversight.

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

VDI AND THE IDENTITY AWARE NETWORK

Many speculated that Virtual Desktop Infrastructure (VDI) was just going to be a phase, but on the contrary, VDI has not only proven to be a contender, but has become a growing necessity for IT departments. Although it may appear that interest in VDI has gone into a downward spiral, this enabling technology, according to a certain number of analysts, is actually on the rebound.

Historically, data centers have been designed for physical environments. This means that companies today are applying yesterday's tools and capabilities to address current and future opportunities: trying to virtualize IT architectures that were not necessarily optimized for virtualization, especially in terms of the network. VDI, in particular, has and will continue to become a staple in the new data center and crop up in progressive IT departments at colleges, law firms, hotels and retail establishments. The business and technical efficiencies involved with establishing VDI are relatively simple and straightforward considering the significant improvements VDI can deliver to network manageability, security and energy efficiency.

identity management are critical network security considerations as users can connect to the data center from any location, using a variety of devices. At the network level, more granular network access policies based upon user roles, device types and physical locations are required. The network then has to scale bandwidth, manage converged communications and implement network layer security policy independently from any single device or application. The access management and lack of identity features of old networks, however, won't be up to par. Before moving to VDI, data center managers must understand how network performance will be impacted while maintaining cost, delivery of multiple converged services and power efficiency.

"VDI delivers significant improvements to network manageability, security and energy efficiency.

According to The Cloud-Based Virtual Desktop Infrastructure Market 2012-2017 published in September 2012 by VisionGain, the VDI market was expected to grow to $11.2 billion by the end of 2012 and continue to increase at a compound annual growth rate (CAGR) of 14.77% through 2015. Large enterprises are drawn to VDI because of its ability to reduce desktop support and management costs, as well as lower overall energy requirements of virtual desktops.

To meet compliance and security regulations, VDI also provides business continuity and disaster recovery capabilities within the data center. VDI does, however, demand an identity approach and a more aware network. Policy and

How will VDI help increase energy efficiency?

While VDI is partially driven by the use of lighter-weight devices, such as smartphones and tablets, the network plays a key role. It helps to reduce energy consumption through the centralization of resources and by bringing much higher speeds at the port level. VDI allows higher density 10 Gigabit Ethernet (GbE) port modules on chassis type switches providing the advantage to collapse all traffic easily into just a few network switches. This is all possible through the consolidation of horsepower into a single core layer as opposed to deploying distributed GbE LANs and multiple tiers, thus providing the necessary bandwidth for all VDI connections. Over-

VDI has shown itself to be more powerful and easier for IT departments to manage, while also achieving the goal of being "green" and highly efficient.

all, VDI has shown itself to be more powerful and easier for IT departments to manage, while also achieving many businesses' goals of being "green" and highly efficient.

How can the network converge voice, video and data for VDI deployments?

After addressing the considerations for system centralization and bandwidth, IT departments will be faced with the next issue of how to carry converged media – mixed voice, video and data. To deliver voice and video traffic to users on predetermined priorities, it is imperative that the network be capable of not only 10 Gigabit and Gigabit to the Edge, but also intelligence, Quality of Service (QoS) and ultra-low latency switching. Just like traditional networks, for users to be able to experience consistent and predictable interactions, the backbone of a VDI network must be capable of handling convergence flawlessly. Once the network has been deemed seamless and demonstrates the required quality, it is at this point that critical activities like IP phone calls and collaboration, e-learning activities using IP video and call centers can be functional.

How does the network support the security of VDI deployments?

With the VDI network, traditional operating systems are eliminated, yet user log on, security policies, visibility and monitoring are required more than ever. Until today, many companies have relied upon the security of traditional networks that entailed complex "application layer" elements of sign-on security, including strong authentication, Single Sign-On (SSO) systems and LDAP directories. For the companies that have taken the leap toward the emergence of VDI, security, including network identity, is now simplified, centralized and managed by the network instead of a PC OS.

With the growing number of secure government facilities choosing to use advanced identity management, which is only possible via VDI, we can expect the private sector to follow by enabling similar deployments for their mobile workforces. To this end, today's businesses looking to deploy VDI securely require a new model called identity-aware networking – a term Jon Oltsik, principal analyst of Enterprise Strategy Group at Extreme Networks defines as "*a policy-based network architecture that understands and acts upon the identity and location of users and devices.*"

With identity-aware networking, the network gathers information from multiple existing sources in an integration process enabling IT departments and managers to use this data to build and enforce access policies. With this rich data (i.e., user, device, and location) in hand, network administrators can easily configure extremely granular network access policies that can then be enforced based upon any or all of this information and/or other factors. For example, the CFO of a company could be granted access to the end-of-quarter financials from their laptop on the corporate LAN or on a home computer connected over a VPN, but not from other devices or networks. In other cases, a contractor may be able to access engineering plans during specified hours only.

In this situation, user and device location is important for several reasons. For starters, identity-aware networking confirms if the user is logging on from a trusted or untrusted network. This location awareness may also be important

In addition to who is connecting to the network, identity-aware networking verifies what is connecting to the network...

depending upon whether a user is accessing the network from a wired port or over Wi-Fi. Furthermore, depending on the device location, whether within a single facility or between two facilities on the same campus, the network access policies may change. In addition to who is connecting to the network, identity-aware networking verifies what is connecting to the network. This is important as different devices (i.e., laptops, tablets, smartphones, power meters, etc.) have different security and performance characteristics. Just as a contractor may only have access during certain times of the day, their laptop may also be treated differently than a remote employee's PC that meets all corporate security and configuration policies.

Network-based identity is associated with information including IP and MAC addresses, VLAN tags and subnets which can all play a significant role in device authentication, VPNs and IPSEC – thus network layer security takes over. The network-based identity looks at a number of specifics when securing the network ranging from the number of inputs (user ID and the role of the user), device characteristics and capabilities which are specific to that unique device, and user/device location.

With most deployments, IT managers and departments strive to meet daily challenges and varying mobile users and devices. To make this possible, more granular network access policies based upon user roles, device types, and physical locations will be required at the network level – a situation that can be handled by VDI. Once this has been achieved, the network then has to scale bandwidth, implement network layer security policy independently from any single device or application, and manage converged communication appropriately.

Although virtualization may have started out as a technology driven by server consolidation, today's evolving network takes it well beyond servers to a means of centralizing IT itself. The evolution starts with desktop PCs and the new class of wireless computing devices proliferating throughout the enterprise, including smartphones, tablets and PCs. In today's Internet-connected world, large organizations need their networks to enable any user to connect securely to applications and services from any authorized device.

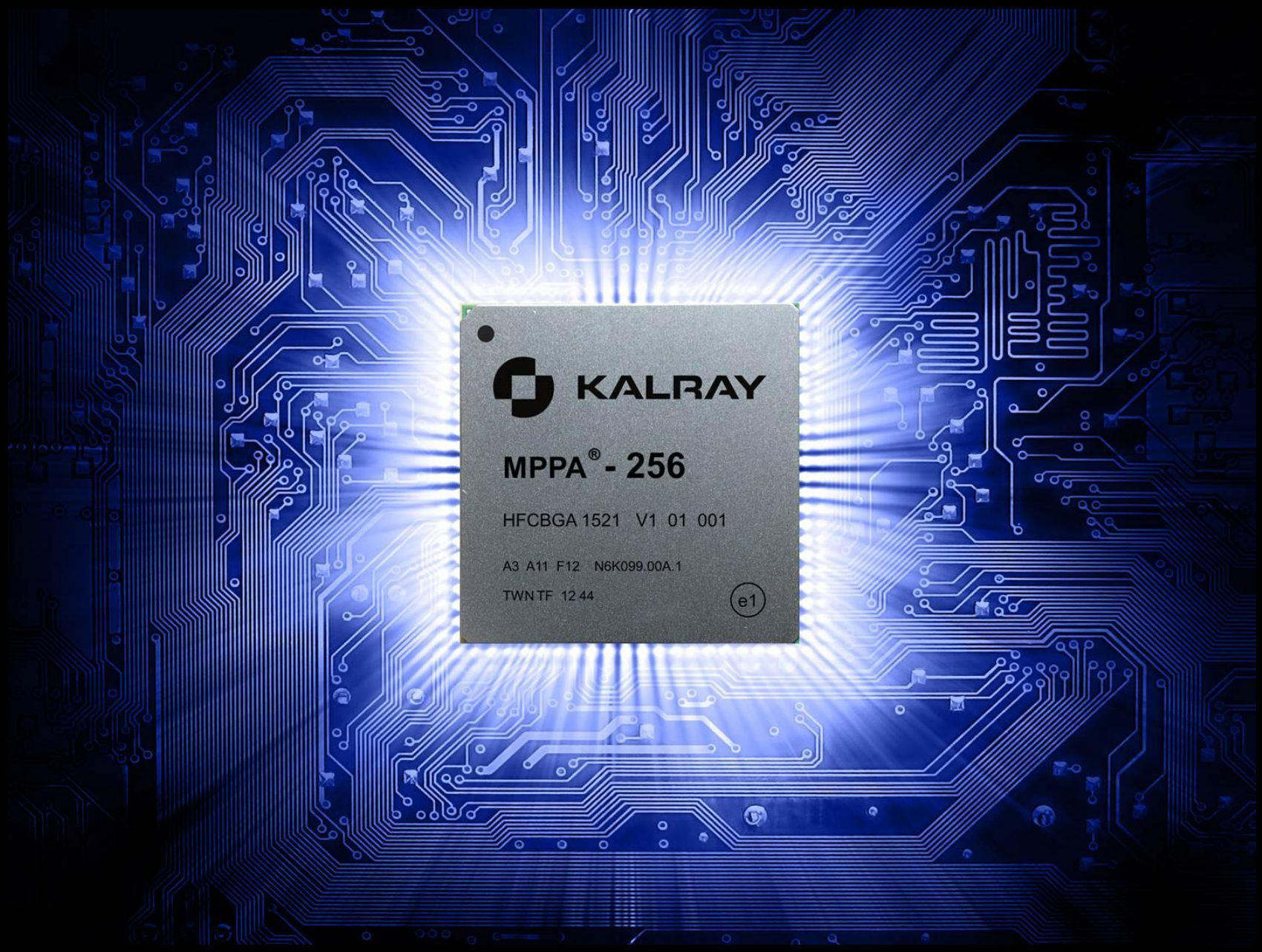
What companies, both enterprises and SMBs, need is a "best of breed" network with the intelligence to not only enforce their policies once users are on the network, but to also dynamically collect information about the users in the network, the devices trying to connect to the network and where the users are physically when trying to connect to the VDI infrastructure. In the end, IT departments will be able to focus more on other tasks such as doing business, regulatory compliance and security ROI benefits because they will dedicate less time to maintaining and managing application-layer security.

VDI demands an identity approach and a more aware network, and only when data center managers closely examine the network's role in meeting key criteria such as cost savings, power efficiency, user and device identity, and ease-of-use can VDI truly progress towards becoming a new norm in computing.

Renuke Mendis
Principal Technical Marketing Engineer
Extreme Networks

KALRAY

AGILE PERFORMANCE

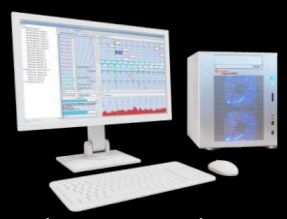


Kalray MPPA®-256

256 cores, 25 GFlops/Watt

World first massively parallel, scalable, MIMD processor optimized for high performance computing and power consumption

www.kalray.eu

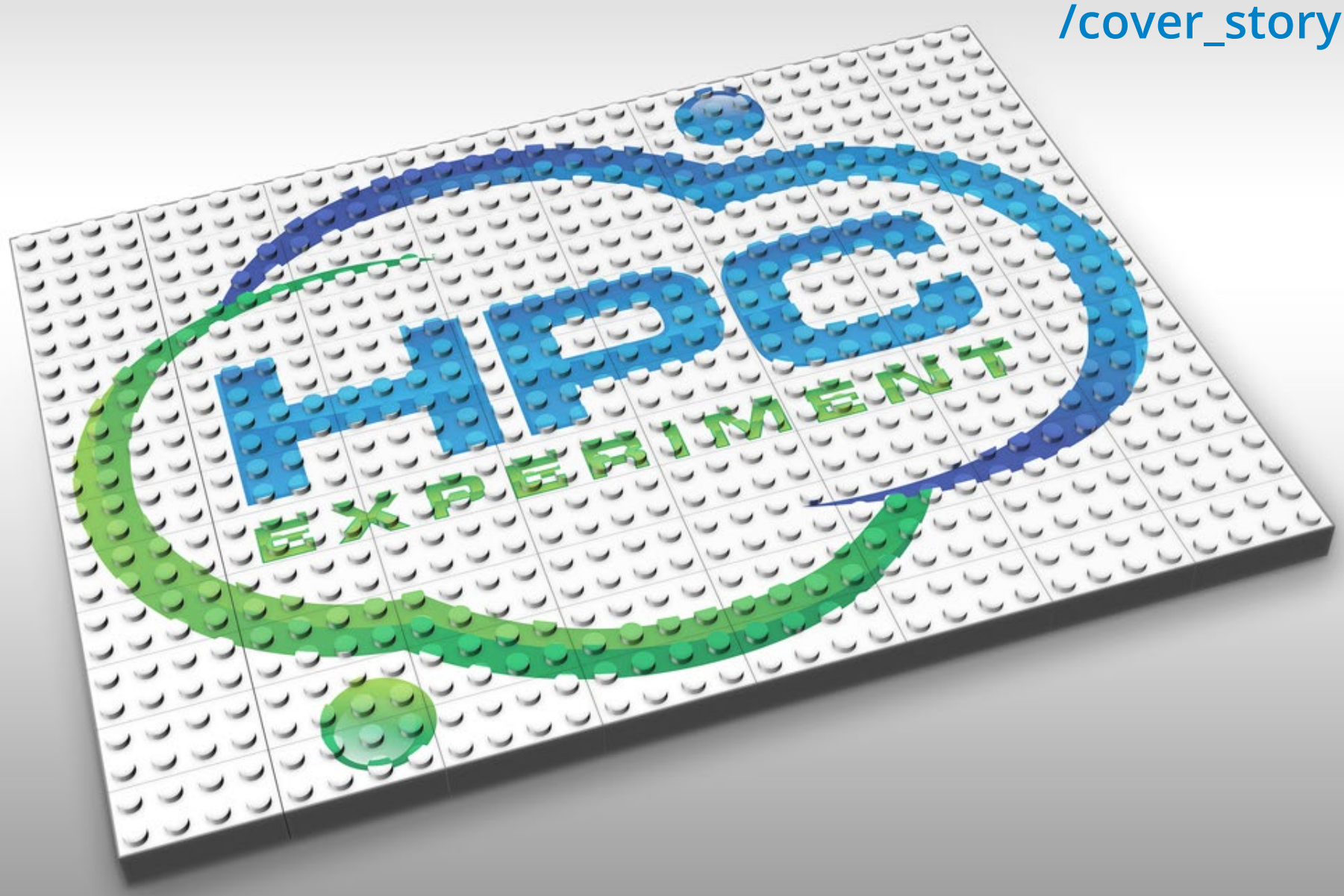


Development workstation
MPPA DEVELOPER

SDK C/C++/FORTRAN, debugger and trace
MPPA ACCESSCORE

Reference boards
MPPA BOARD

/cover_story



AN INSIDE LOOK AT

THE UBERCLOUD HPC EXPERIMENT

Once a matter of theoretical debate, high performance computing in the cloud is now becoming an actionable reality. After reviewing (and demystifying) the issues traditionally associated with it - performance, cost, software licensing, security - let's take an insider look at the advances, challenges, and a few application use cases from the UberCloud HPC Experiment...

WOLFGANG GENTZSCH* / BURAK YENER*

Cost savings, shorter time to market, better quality, fewer product failures... The benefits that engineers and scientists could achieve from using HPC in their research, design and development processes can be huge. Nevertheless, according to two studies conducted by the US Council of Competitiveness (*Reflect* and *Reveal*) [1], only about 10% of manufacturers currently use HPC servers when designing and developing their products. The vast majority (over 90%) of companies still perform virtual prototyping or large-scale data modeling on workstations or laptops. It is therefore not surprising that many of them (57%) face application problems due to the inadequacy of their equipment; more precise geometry or physics, for instance, require much more memory than a desktop could possibly possess. Today, there are two realistic options to acquire additional HPC computing: buy a server or use a cloud solution.

* Co-founders, The UberCloud.

Many HPC vendors have developed a complete set of HPC products, solutions and services, which makes buying an HPC server no longer out of reach for an SME. Owning one's own HPC server, however, is not necessarily the best idea in terms of cost-efficiency , because the Total Cost of Ownership (TCO) is pretty high, especially considering that maintaining such a system requires additional specialized manpower.

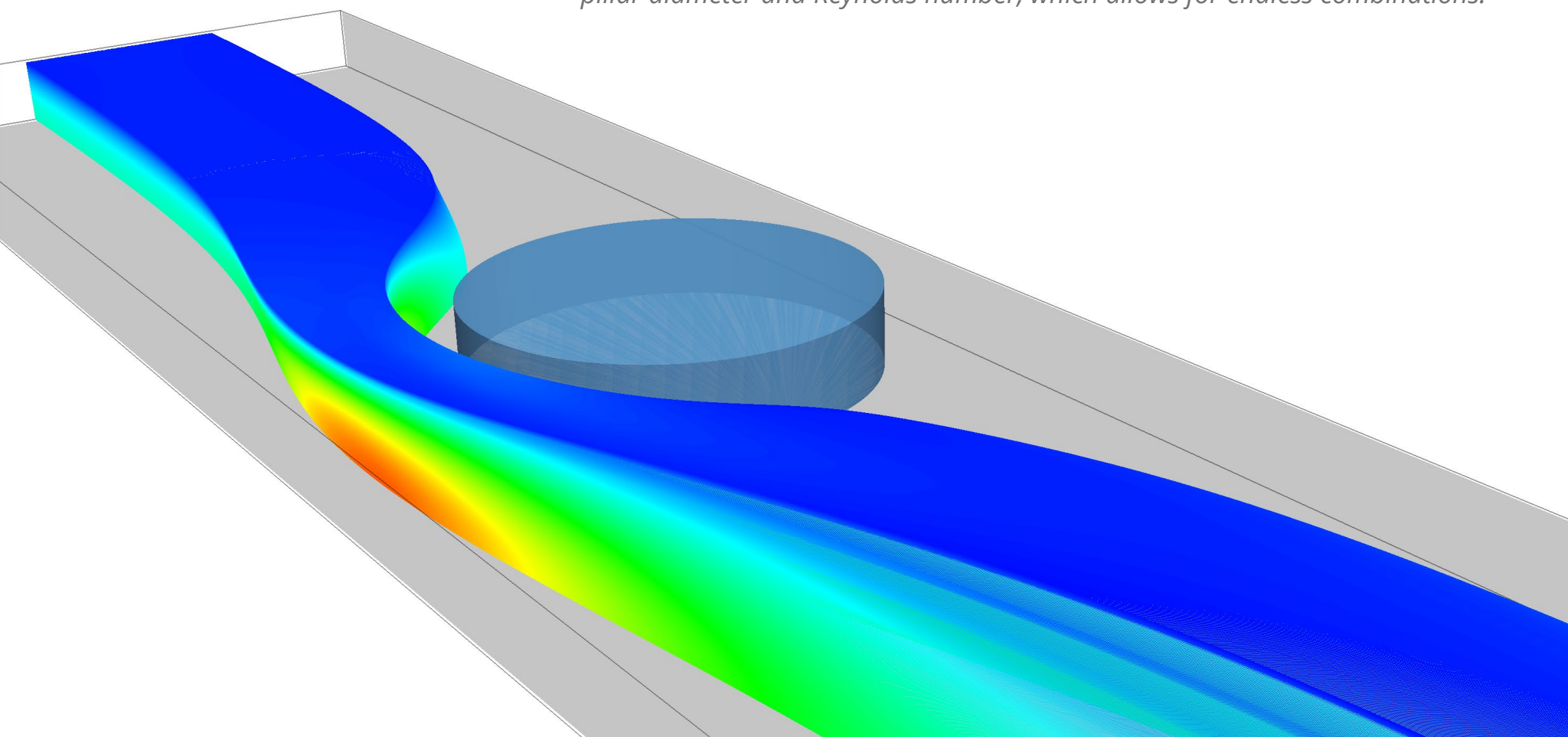
In addition to the high costs of expertise, equipment, maintenance, software and training, buying an HPC system also often requires a long and painful internal procurement and approval processes. The other option is to use a cloud solution that allows engineers and scientists to continue using their regular computer system for their daily design and development work and to " burst" the larger, more complex jobs into the HPC cloud as needed. In this way, users have access to virtually limitless HPC resources that offer higher quality results. In management and financial terms, a Cloud solution helps reduce capital expenditure (CAPEX). It offers businesses greater agility by dynamically scaling resources as needed and is only paid for when used.

1 - What is an HPC cloud?

According to the National Institute of Standards and Technology (NIST) [3], "*Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.*" NIST further explains that the cloud model is composed of "five essential characteristics" (on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service), three "service models" (Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) and four "deployment models" (private cloud, community cloud, public cloud and hybrid cloud).

Standard cloud services can address a certain portion of HPC needs, notably those that don't require a lot of parallel processing such as parameter studies with varying input and low I/O requirements traffic. However, many HPC applications cannot be shoehorned into standard cloud solutions and consequently require hardware designs that can efficiently run certain HPC workloads from various science and engineering application areas.

Team 53 - Understanding fluid flow in microchannels with the insertion of a pillar. Four variables characterize the simulation: channel height, pillar location, pillar diameter and Reynolds number, which allows for endless combinations.



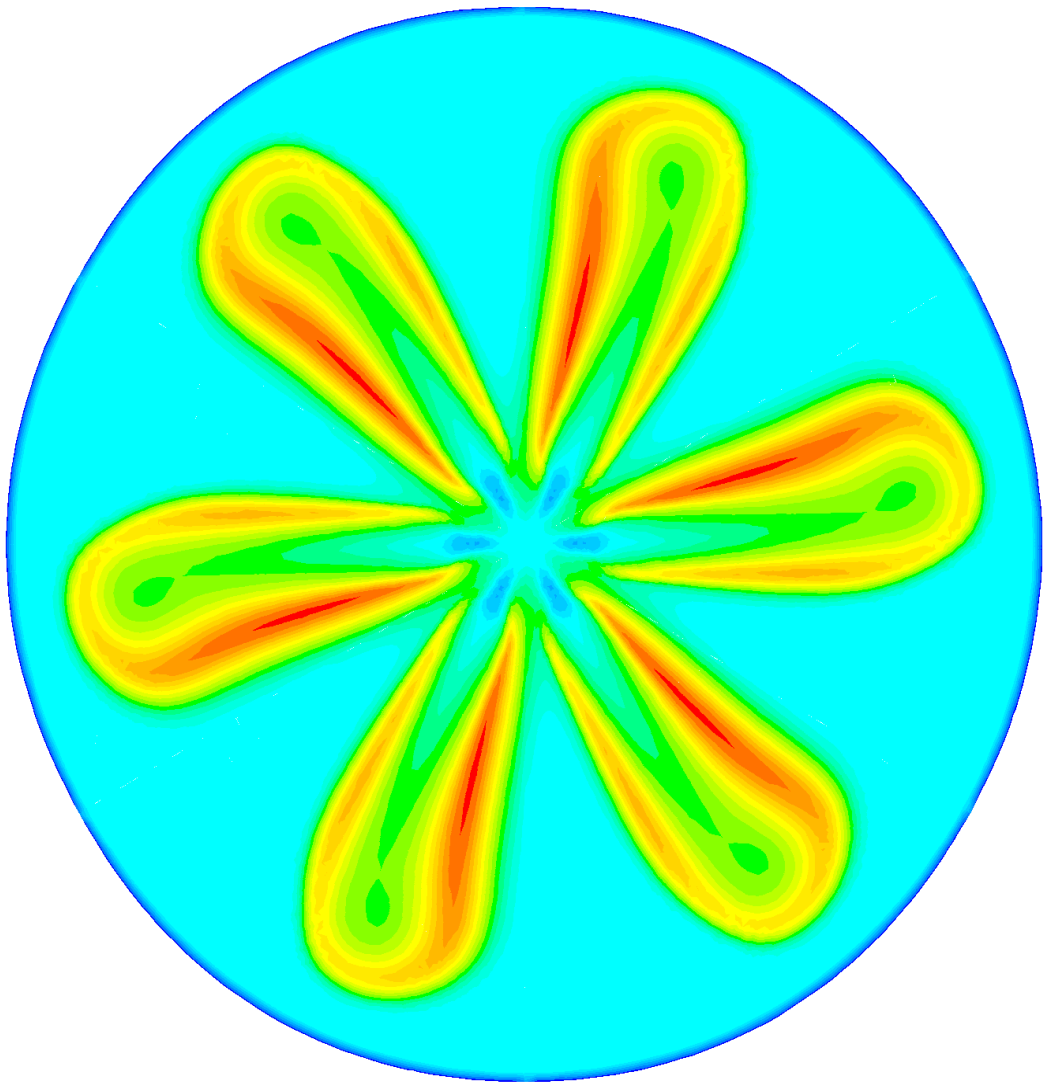
Many distributed computing applications are developed and optimized for specific HPC systems and require intensive communication between parallel tasks. To perform efficiently in an HPC cloud, these applications require additional system features, such as:

• Large capacity and capability resources, application software choice and a physical or virtualized environment depending on performance needs. The use of high performance interconnects and dynamic provisioning can offer cloud features while maintaining HPC performance levels

• High performance I/O is often necessary to ensure that many I/O-heavy HPC applications will run to their fullest potential. For example, pNFS might provide a good plug-and-play interface for many of these applications. Back-end storage design, however, will be crucial in achieving acceptable performance.

• Fast network connection between the high performance cloud resources and the end-user's desktop system. Scientific and engineering simulation results often range from many Gigabytes to a few Terabytes. Additional solutions in this case are remote visualization, data compression, or even overnight express mailing a disk with the resulting data back to the end-user (by the way, a quite secure solution).

Additionally, there may be other issues that need to be addressed before an HPC cloud can deliver low-cost and flexible HPC cycles; a careful analysis of application requirements is required in order to determine effective HPC performance in standard cloud offerings.



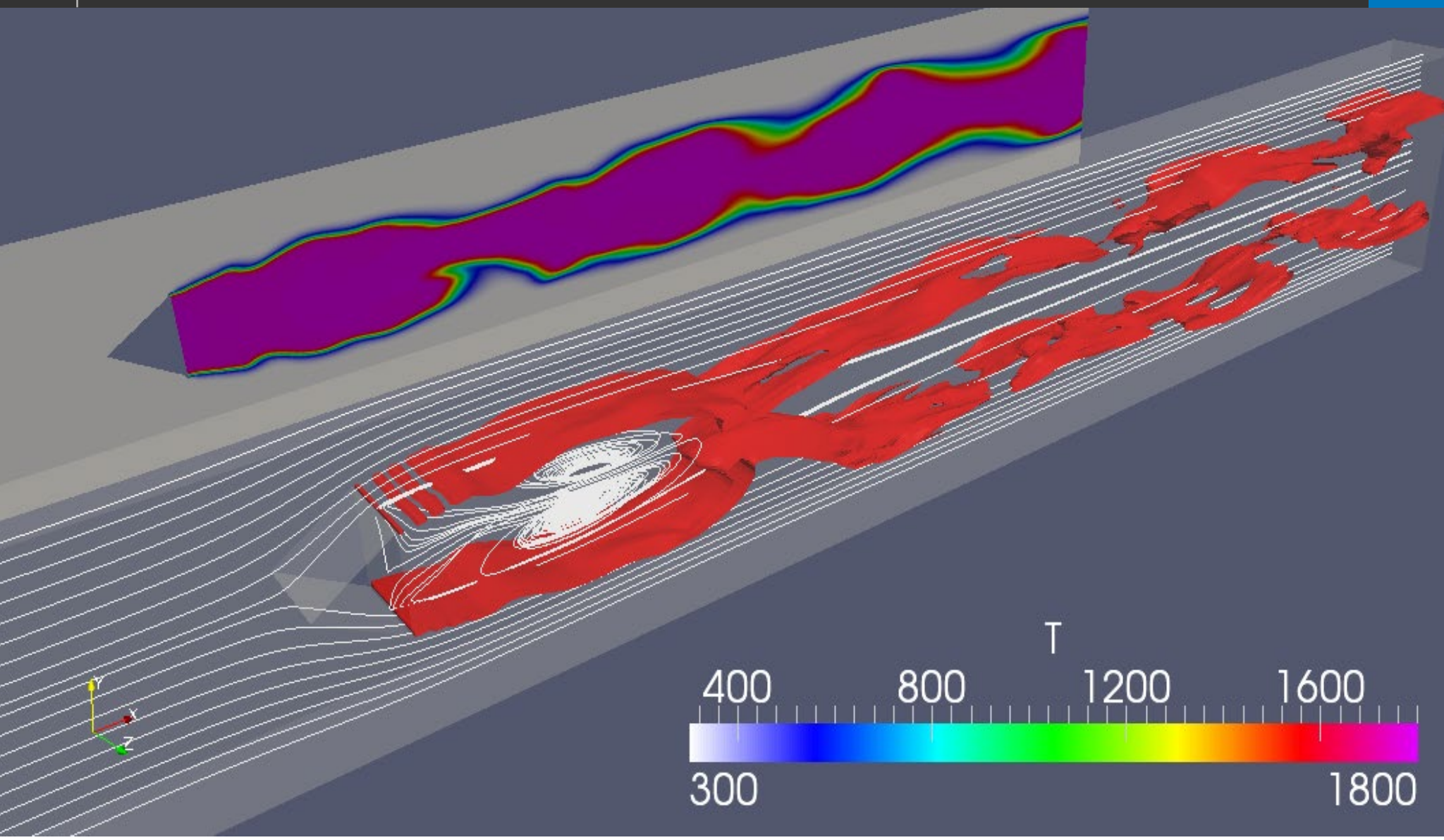
Team 36 - Advanced combustion modeling for Diesel engines. Simulation result showing the flame (red) located on top of the evaporating fuel spray (light blue in the center).

2 - Security in the cloud

Security is the major issue in all cloud deployments and this concerns every organization, not just SMEs. In order to protect its "crown jewels", any organization migrating to the cloud should primarily be concerned with the physical location of its data.

It should also address a number of questions related to back-up and recovery, auditing, certification and of course, security. In particular, who on the service provider's side will have access to the data. Achieving acceptable levels of security is not only a matter of technology, but also a matter of trust, visibility, control and compliance. This is achieved through effective management and operational practices. In other words, security is first and foremost a human issue, because when end-users send data and applications into the cloud, they sacrifice complete control.

However, security does not appear to be a major issue for cloud customers, as the market is currently growing by approximately 30% per year. Optimistically, there is no reason to believe that HPC in the cloud will not follow the same trend.



Team 52 - High-resolution simulations of blow-off in combustion systems. Predicted temperature contours field using OpenFOAM.

We could not agree more with Simon Aspinall [4] when he says *"as with any other new and disruptive technology, there is still some hesitation and some assumptions about how businesses should run, that have until now inhibited the speed of adoption in the enterprise market. These arguments, especially around security, are oddly similar to those that were once voiced around the Internet, online retail and even the mobile phone. Businesses not doing so already would be well advised to take a page from history and adapt or risk getting left behind by competitors already benefiting from the efficiencies a Cloud solution delivers. As with any other major decision, the key is to educate ourselves on the available options, do diligence, define our "Cloud," and adapt at a pace that is best for us."*

3 - The HPC cloud market, providers and applications

While the general enterprise cloud services market is currently reaching 5% of the overall IT spending worldwide, HPC in the cloud is still somewhat behind. The only market data currently publicly available for HPC clouds was produced by IDC for a worldwide study involving 905 HPC sites of various sizes in government, academia and industry/commerce [5], a summary of which was presented at the ISC Cloud Conference in September 2013 in Heidelberg.

Among other interesting data, the study showed that 23.5% of these sites used cloud computing of some type in 2013, a substantial increase from 13.8% in 2011. Of the HPC sites that use cloud computing, half use private clouds and slightly more than half use public clouds (probably at an early stage). Early adopters include: government, manufacturers, bio-life science, oil and gas, digital content creation, financial services and other high-end analytics, online consumer services and health care firms. Our own observation is that the majority of these early adopters are using cloud computing mainly for exploring this new paradigm and how it might fit into their current strategy. Only a small percentage of them are actually using cloud in production.

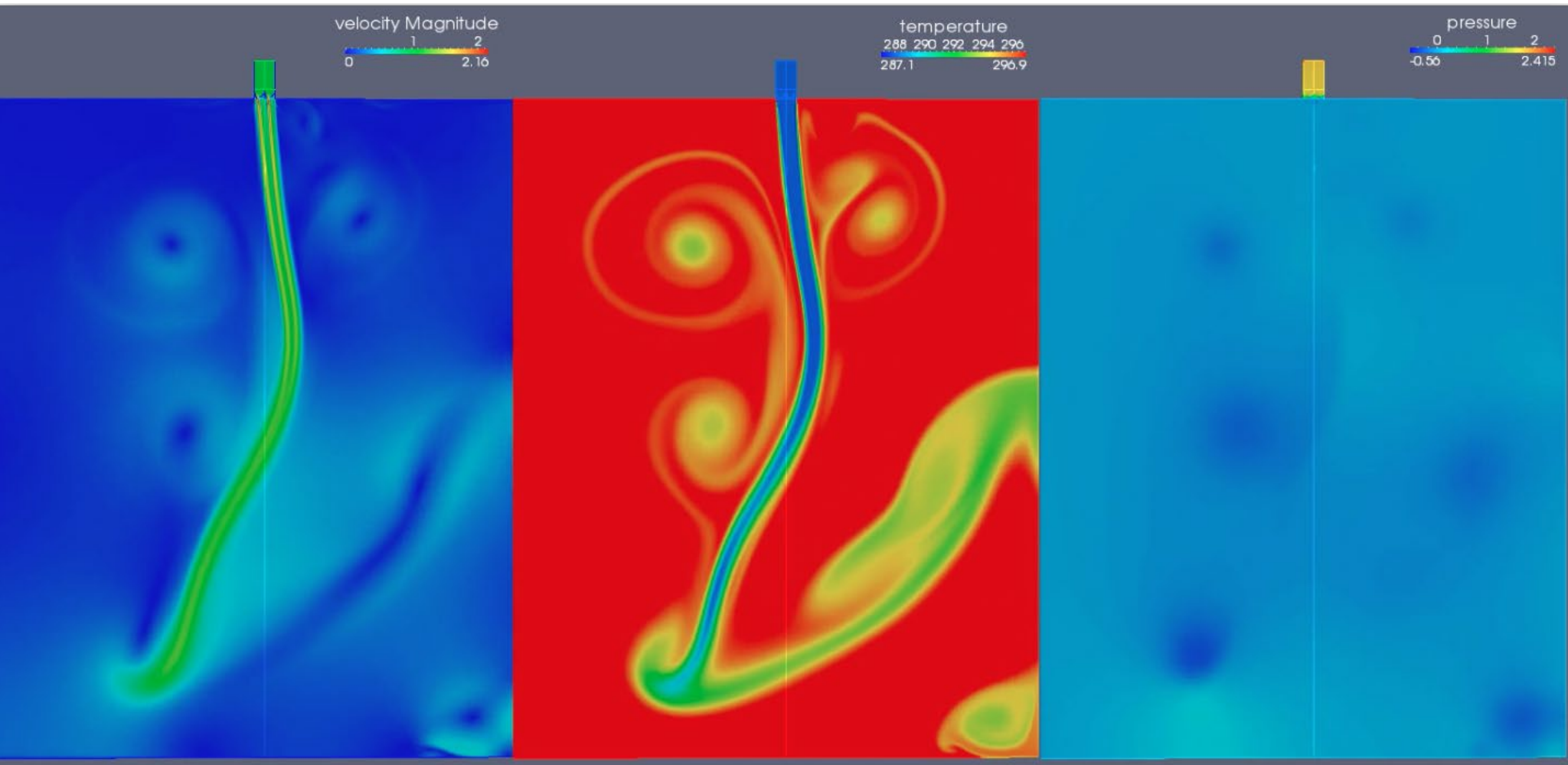
3.A - HPC cloud adoption for SMEs

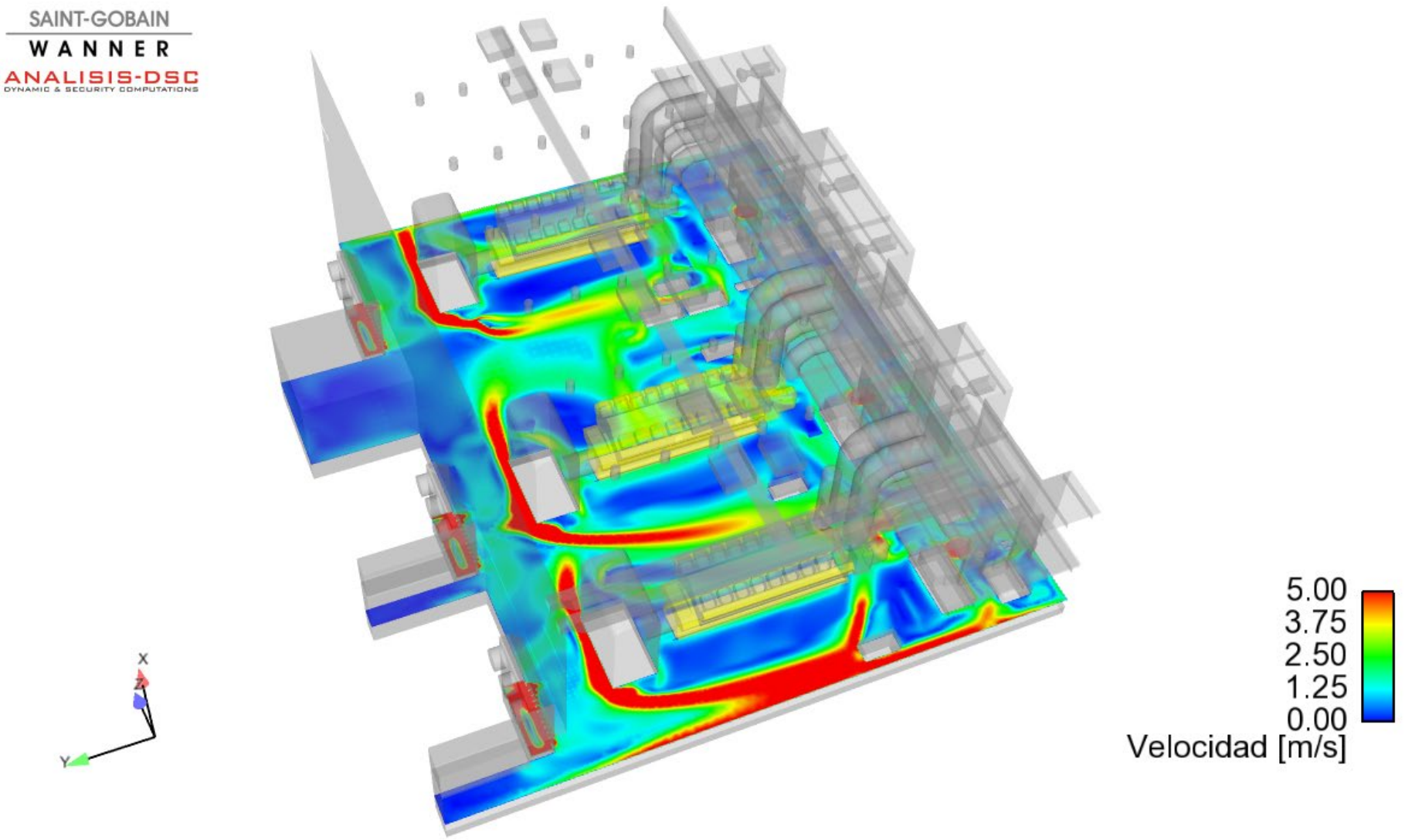
In their study, IDC did not take into account one significant market segment: digital manufacturing, in particular those only using desktop workstations. This market is mainly composed of engineers and scientists who need advanced computing technologies for simulations and production in order to improve quality, achieve faster time-to-market and reduce costs. They can be found in a multitude of areas, such as structural analysis, aerodynamics, fluid flows, crashworthiness, environment testing, stress testing, process engineering and manufacturability. As previously noted, they primarily use workstations, but according to Jon Peddie [6], the worldwide workstation market is currently worth $7 billion in terms of revenue, with around 3.5 million units shipped last year. That's easily in the neighborhood of 20 million users, mainly using workstations and PCs for their daily design and development simulations, making them potential candidates for HPC clouds, let alone those manufacturers who don't use computing at all (sometimes called the 'missing middle') or even the knowledgeable amateur who could do a lot with sufficient computing power, in the context of 3D printing for example.

A confirmation of this potential can be found in an Intersect360/NCMS study on Modeling and Simulation at 260 U.S. Manufacturers [7], which we believe is still valid today. This study shows that in 2010, 61% of companies with over 10,000 employees were using HPC, whereas only 8% of companies with fewer than 100 employees were.

Meanwhile, at the same time, 72% of desktop CAE users saw a competitive advantage in adopting more advanced computation. Obviously, there are a few compelling roadblocks towards adopting HPC, especially in the mid-market. As previously mentioned, these roadblocks are principally the TCO and the cumbersome procurement processes. Clearly, because of the low penetration of HPC beyond the desktop in the general manufacturing industry, the use of HPC in the cloud is still in its infancy.

Team 30 - Heat Transfer Use Case. 2D room with a cold air inlet on the roof, a warm section on the floor and an outlet on a lateral wall near the floor. Picture shows the velocity and temperature fields at a certain time of the transient simulation.





Team 54 - Analysis of a pool in a desalinization plant with complex air-water free-surface modeling.

And yet, there are rising hopes that this situation will change. First and foremost, because the technical press regularly addresses the various roadblocks and helps raise awareness of the many benefits of in-house HPC and HPC as a Service (HPCaaS) in the SME market of the manufacturing industry. Secondly, because there are several important trends and initiatives that foster HPC in the cloud for this same market. Thirdly, because there is a growing acceptance of general enterprise cloud computing solutions such as ERPs or CRMs, which will eventually contribute to the adoption of HPC-related services for engineering simulations on-demand.

Additionally, large manufacturing companies expect their supply chain partners to perform more and more high-quality end-to-end simulations in less time on HPC systems. And, last but not least, there are american and international initiatives like the 'Missing Middle', the National Center for Manufacturing Sciences, the UberCloud HPC Experiment that help promote the concept of HPC in the cloud.

3.B - A growing number of HPC cloud service providers

Over the past five years, hundreds of cloud services providers (CSPs) have entered the market, offering hardware resources, software and expertise. To name just a few resource providers: Amazon AWS, CloudSigma, Fujitsu TC Cloud, GOMPUTE, GreenButton, HP, MEGWARE, Microsoft Azure, Nimbix, OCF, Oxalya/OVH, Penguin Computing, Rescale, Sabalcore, Serviware/Bull, SGI, SICOS, TotalCAE and transtec (more providers here).

While commercial CSPs evolve, a growing number of supercomputing centers are becoming interested in offering access to their HPC clusters expertise to local and/or regional industry players on a pay-per-use basis. Examples include Georgia Tech, NCSA, Ohio Supercomputing Center, Rutgers University the San Diego Supercomputer Center in the US, the CESGA Supercomputing Centre and FCSCL in Spain, CILEA in Italy, GRNET in Greece, HSR in Switzerland, Monash University in Australia, SARA in The Netherlands and the Mesocentres initiative in France.

In the late 90s, many digital manufacturing ISVs tried to get into the Application Service Provider business by offering access to their solutions via the web. Most of them failed because of severe (or rather insurmountable) technical or busi-

ness roadblocks that couldn't be overcome at that time. While many of the barriers have been resolved or at least softened, in the meantime many of these ISVs remain concerned that the cloud computing paradigm could potentially disrupt their existing business model based on selling annual node-locked or floating licenses. The majority of them still believe that customers might turn to on-demand cloud-based pay-per-use software services instead of buying and using software on their own systems.

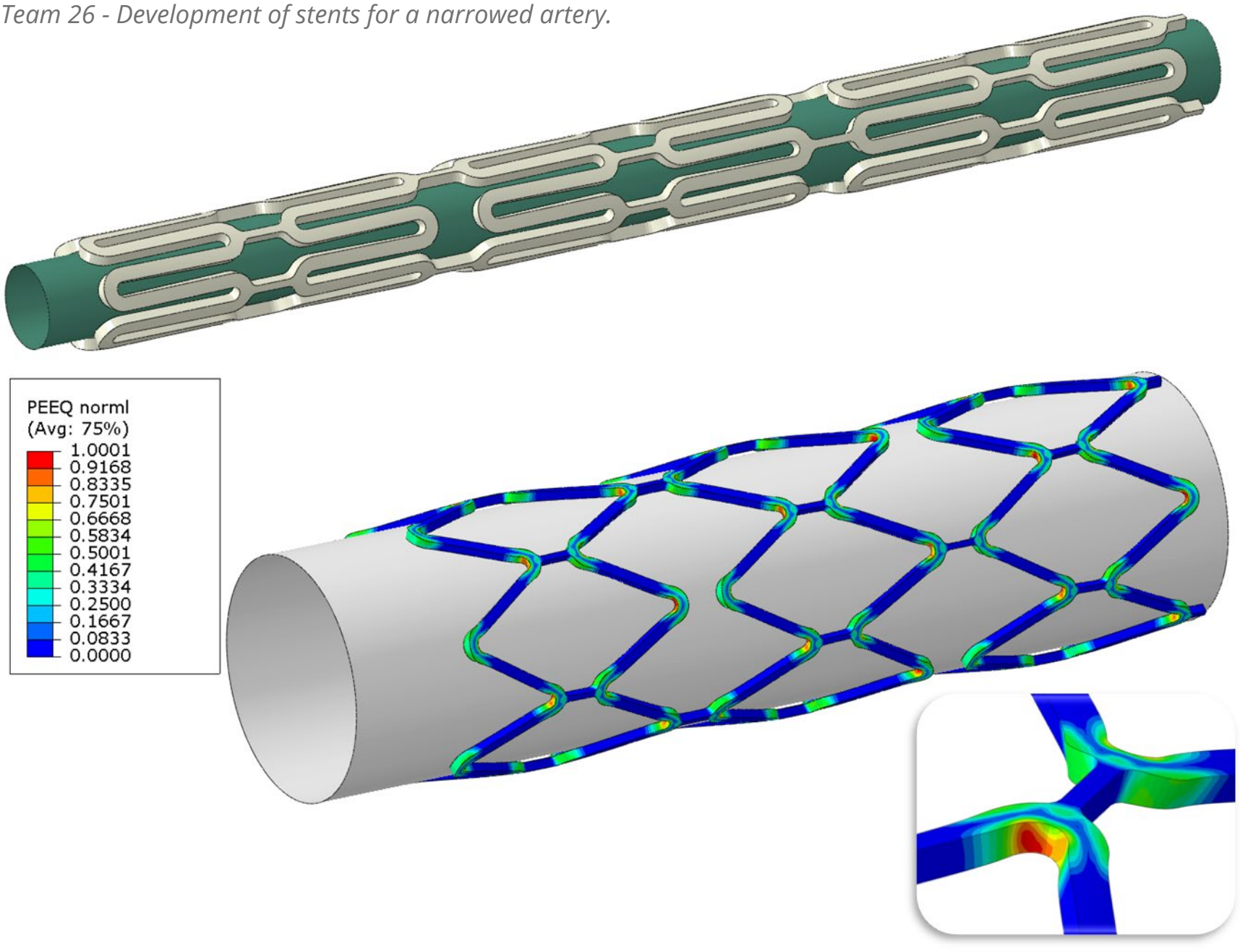
While this apprehension is understandable, the reality may be totally the opposite... In fact, engineers who turn to the cloud because of the above-mentioned benefits are very likely to also continue using their workstations for day-to-day standard R&D scenarios.

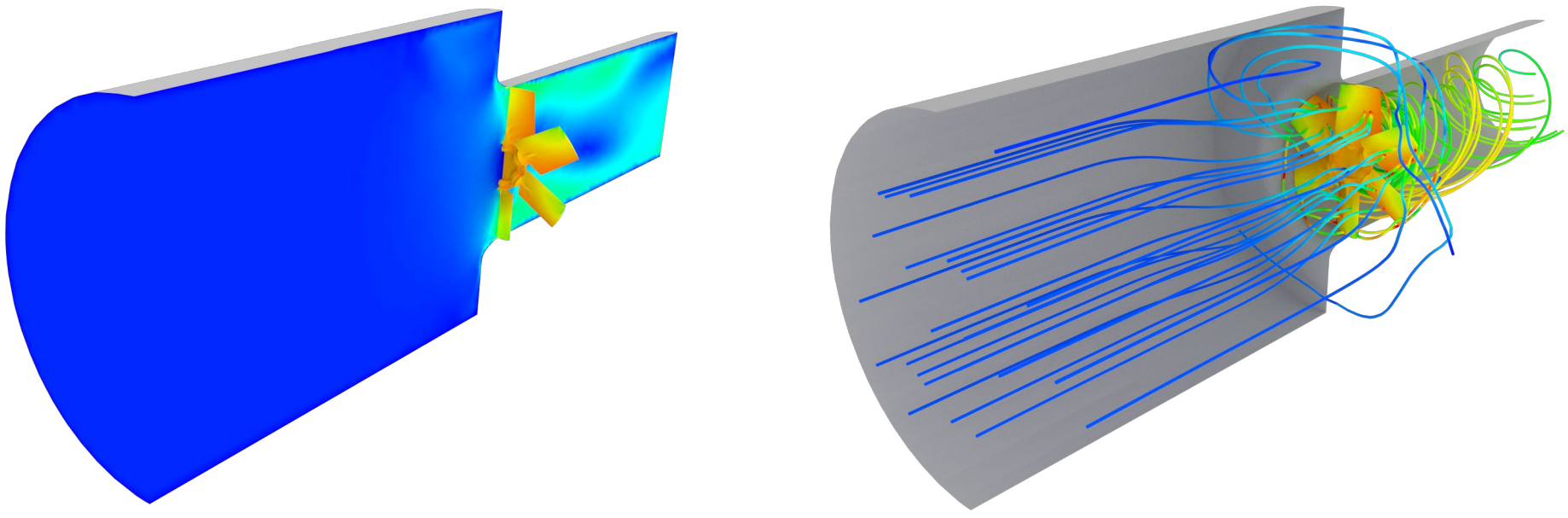
The cloud offers increased opportunities for engineers to do more, faster and better simulations (and products) - see the example in [8]. For these scenarios, ISVs will sell on-demand, pay-per-use, hourly rated or subscription tokens, resulting in additional business as well as workstation license revenues. Even for those companies which already have an HPC cluster in their computing center, bursting into a cloud offers higher efficiency and flexibility for the engineering team and for the computing center, which will more likely result in increased on-demand license business. Not to mention that ISVs offering on-demand application software will be able to attract new customers who are just beginning to work with computer simulations and who would have never considered buying a license for just a few simulations.

3.C - HPC cloud software licensing

HPC cloud software licensing is still considered one of the major barriers for the wider adoption of cloud computing. This is particularly true for SME manufacturers, as shown by a poll during an UberCloud Webinar in June 2013: the ISVs' slow adoption of more flexible (on-de-

Team 26 - Development of stents for a narrowed artery.





Team 56 - Simulating the performance of an axial fan in a duct. Typical results (speedup over number of cores) of a simulation on a workstation (12 cores, green dot), in-house HPC Cluster (16 and 24 cores, blue line) and HPC Cloud (12, 24 and 48 cores, red line).

mand) licensing models for the cloud was the major concern for 61% of the respondents. At the same time, things are a-changing. On the one hand, the very same ISVs that have been traditionally reluctant to provide on-demand licenses are now seriously considering them (ANSYS, SIMULIA…) or are already offering SaaS services (Autodesk Sim360, CD-adapco's Power on Demand…). And on the other hand, a large number of major software providers - especially the ones involved in digital manufacturing or HPC tools with cloud features - are currently participating in the UberCloud HPC Experiment.

Among them: Acellera, Adaptive Computing, Advanced Cluster Systems, Advection Technologies, AMPS Technologies, ANSYS, Artes Calculi, Autodesk, BGI, BlackDog Endeavors, Bright Computing, CAELinux, CEI, Certara, CHAM, Ciespace, CloudioSphere, Cloudsoft, Cloudyn, eXact, CPUsage, Cycle Computing, Datadvance, ELEKS, Equalis, ESI Group, ESTECO, Expert Engineering Solutions, Fidesys, Flow Science, Foldyne Research, Friendship Systems, Gompute, GPU Systems, HCL Infoystems, HPC Solves, HPC Sphere, Kitware, Kuava, Landmark (Halliburton), MapR Technologies, micrOcost, migenius, MSC Software, Nice Software, Nimbus Informatics, Numerate, Open Source Research Institute, Ozen Engineering, Personal Peptides, Phenosystems, PlayPinion, Qtility Software, QuantConnect, Rescale, RMC Software, SimScale, SIMULIA, Stillwater Supercomputing, TECIC, TotalSim, TYCHO, Univa and Visual Solutions…

3.D - Application performance in the cloud

The HPC application spectrum is large. On one end, there are the so-called massively (or 'pleasingly') parallel applications like parameter studies in digital manufacturing and drug design. In these cases, the code runs many instances in parallel on many cores, with each parameter job running on one core; or, more globally, applications run on many servers, with each parameter job running on separate servers (with moderate parallelism on the servers' cores). These applications are well-suited for standard enterprise cloud servers.

On the other end of the spectrum, parallel (distributed) applications have tightly-coupled communication needs among parallel tasks and/or high scalability requirements that would run perfectly on specialized HPC servers. A third class of codes has mid-size, mid-scale parallel requirements that can easily run on one server's parallel cores.

Three components are necessary for a cloud datacenter to provide a single-point of access to this kind of heterogeneous HPC cloud composed of enterprise and HPC servers: a user-friendly portal providing access to the cloud, an intelligent resource manager scheduling the different application jobs to the appropriate systems within the HPC cloud and computing resources, ideally preloaded with the requested codes.

4 - The cost model for in-house *vs* in-cloud high performance computing

According to an IDC study [2], only 7% of the total cost associated with acquiring and operating an HPC system comes from the hardware. A much larger portion comes from the high cost of expertise (staffing), equipment, maintenance, software and training required to run such systems, which explains the high TCO.

4.A - The cost of an HPC system

To estimate what it really costs to own and operate an HPC System, let's look at a realistic scenario. Let's assume that a company needs to run a variety of simulation jobs, most of them mobilizing 32 cores (for example, 10 million finite elements) and some of them with finer-grained geometry mobilizing 256 cores (e.g. 100 million finite elements). The company has to buy a 16-node system, each node having 16 cores, to be able to perform the 32-core as well as the 256-core runs.

Let's say the price of this kind of a system would amount to $70K, or 7% of the TCO. According to IDC, the Total Cost of Ownership of such a system is $1,000K over three years, or **$330K per year for 256 cores**, or $1,289 per core per year, or $0.15 for one core per hour (core hour). Googling "average server utilization" returns average utilization rates between 5% and 20% (we know some are higher, especially in dedicated scientific supercomputer centers). The table below shows the total cost of one core hour for a 16-node (256-core) in-house HPC cluster depending on utilization rates (or "number of busy nodes") of this cluster.

4.B - The cost of HPC in the cloud

To get to the real cost of HPC in the cloud, let's consider the cost of one core per hour which amounts roughly to $0.20 (this is a typical price at AWS for example, if you exclude additional services and application software of course). Thus, the workload for a 20% utilized cluster is equivalent to 256 cores * 24 hours * 365 days * 20% = 448,512 core hours * $0.20... or **$90K all in all**.

4.C - The cost of a hybrid solution

Let's go back to our scenario of a company that needs to run a mix of 32-core and 256-core simulation jobs at an average 20% system utilization. Let's also assume the 32-core workload utilizes a small in-house HPC cluster with two 16-core nodes, to accommodate all of the 32-core jobs. This cluster is just 12.5% of the size of the big 256-core cluster with $1,000K three-year TCO, i.e. $42K per year. With an excellent cluster utilization of say 92%, the core hour amounts to $0.16 on this 2-node in-house cluster. If the remaining 256-core 16-node jobs run in the cloud, for about one month per year (we have to choose one month to achieve the above 20% utilization rate of the full in-house cluster), the core hour is again $0.20, or $37K per month.

The combined expenditure of in-house and in-cloud costs ($42K and $37k, respectively) **results in $79K** per year for the hybrid solution, compared with $90K for a full HPC in the cloud service and $330K for the in-house cluster. If the decision is a matter of cost and nothing else, the hybrid and the cloud solutions surpass the in-house HPC cluster by at least a factor of 3 for a 'standard' 20% usage rate, or less. According to the table above, the cost of the in-cloud solution ($0.20 per core hour) exceeds that of the in-house solution only above an average 75% utilization rate of the latter - a figure typical of academic Supercomputing Centers serving hundreds or thousands of users, but not of average private organizations.

Busy nodes	1	2	3	4	5	6	8	12	16
Utilization, %	6.3	12.5	18.8	25.0	31.3	37.5	50.0	75.0	100
Cost: 1 core / h, $	2.36	1.19	0.79	0.59	0.47	0.40	0.30	0.19	0.15

Table 1 - Cost per HPC node depending on utilization rates. Source: IDC.

5 - The UberCloud HPC Experiment accelerates HPC in the cloud

In theory, Cloud Computing and its emerging technologies such as virtualization, web access platforms and their integrated toolboxes, solution stacks accessible on-demand, automatic cloud bursting capabilities, etc., enable researchers and industries to use additional computing resources in a flexible and affordable on-demand way. Now, the UberCloud HPC Experiment provides a platform for researchers and engineers to explore, learn and understand the end-to-end process of accessing and using HPC clouds, to identify the issues and resolve the roadblocks (more in [9]). The principle is that end-users, software / resource providers and HPC experts collaborate in teams to jointly solve the end-user's application problems in the cloud.

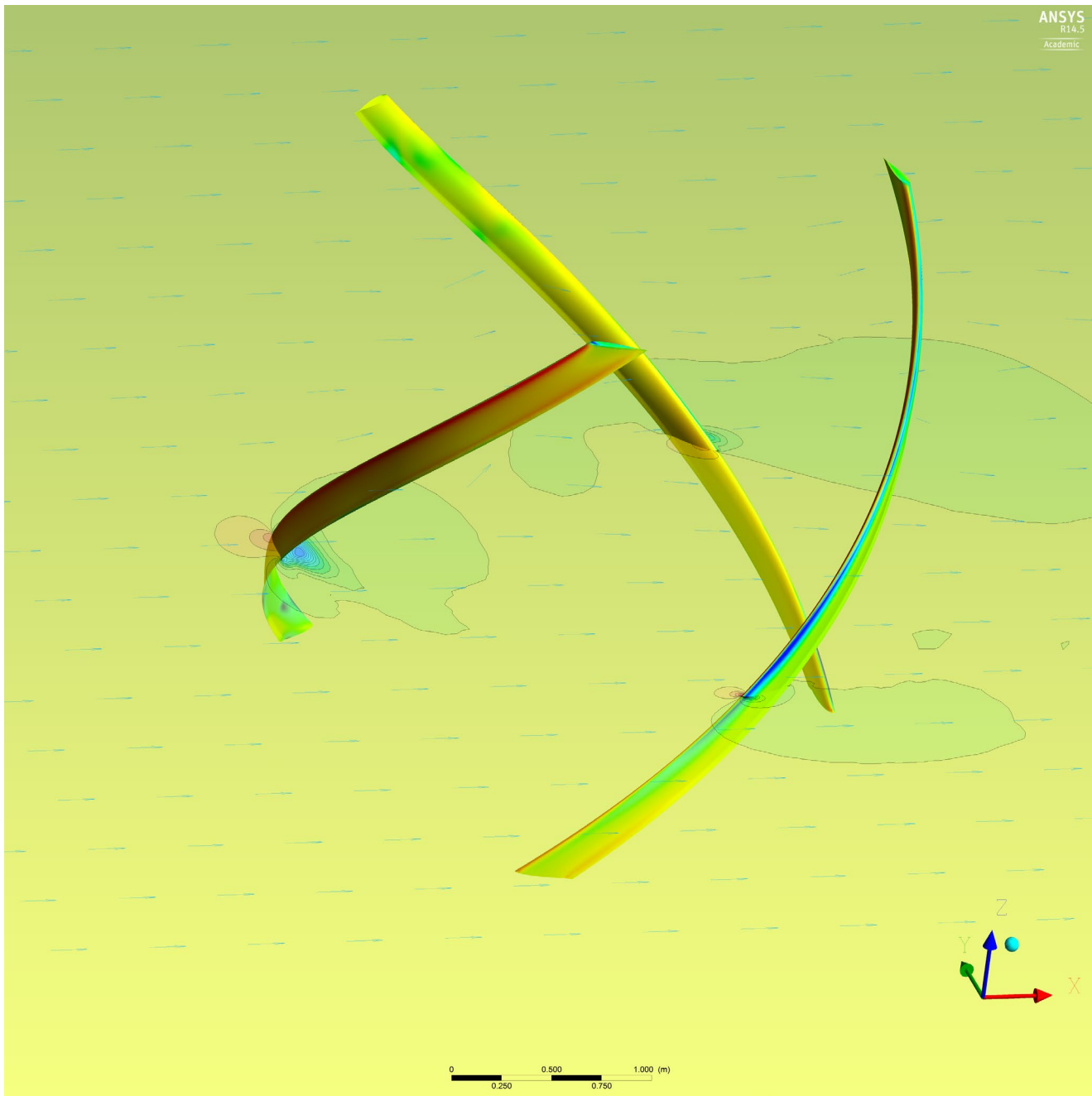
Since July 2012, the UberCloud HPC Experiment has attracted 1,100+ organizations from over 66 countries (as of February 2014). To date, the organizers have been able to build 125 of these teams, in CFD, FEM, Computational Biology and other prominent HPC domains, and to publish more than 60 articles about the UberCloud initiative, including numerous case studies about the different applications and lessons learned. Recently, UberCloud TechTalk and a virtual Exhibition [10] have been added along with a Compendium that includes 25 case studies from digital manufacturing in the Cloud [11].

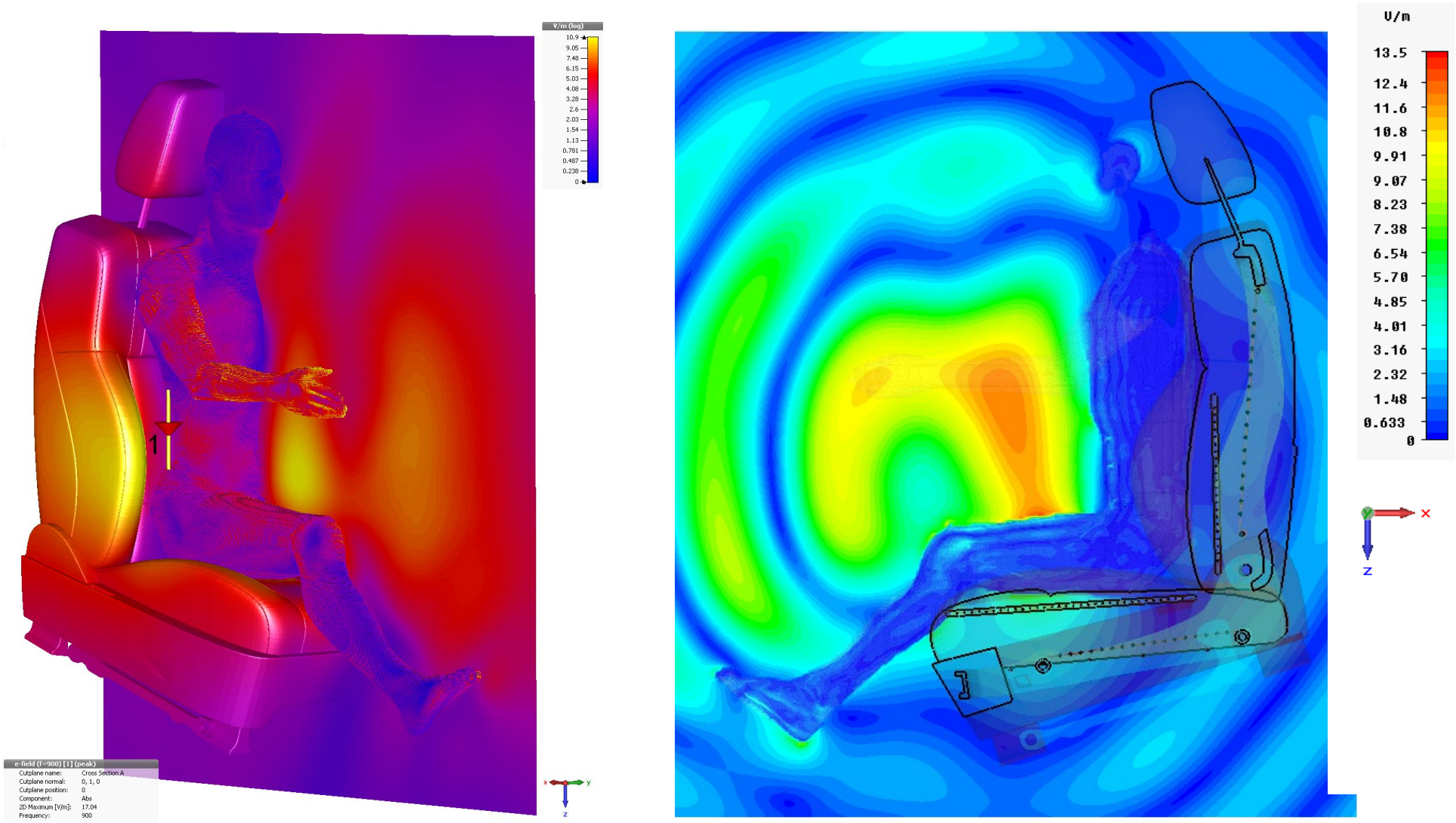
5.A - The UberCloud HPC Experiment History

Inspired by the results of the Magellan Report [12], the UberCloud HPC Experiment idea arose in May 2012 while we were comparing the acceptance of cloud computing for typical enterprises applications in contrast to High Performance Computing.

While the adoption of cloud computing in the enterprise market is rapidly growing (41.3% per year through 2016, according to Gartner [13]), the awareness and adoption of cloud computing in HPC and digital manufacturing is still very slow. This is mainly due to obstacles such as inflexible software licensing, slow data transfer, security of data and applications and a lack of specific architectural features (resulting in reduced performance in the cloud). The idea of the UberCloud HPC Experiment was therefore to find out more about the end-to-end process of bringing engineering applications to the cloud and, in doing so, to learn more about the real difficulties and the best ways to overcome them. The Experiment started in July 2012.

Team 34 - CFD simulation of a vertical wind turbine with 3 helical rotors.





Team 14 - Electric field distribution at 900MHz for 1W excited power displayed on model surface and on a cutplane. The anatomical human voxel model ("Duke") was realistically posed on a car seat in front of a dipole antenna representing a mobile device with hands-free equipment discretized with 112 million mesh cells. Simulation performed with transient solver of CST MICROWAVE STUDIO based on the Finite Integration Technique (FIT).

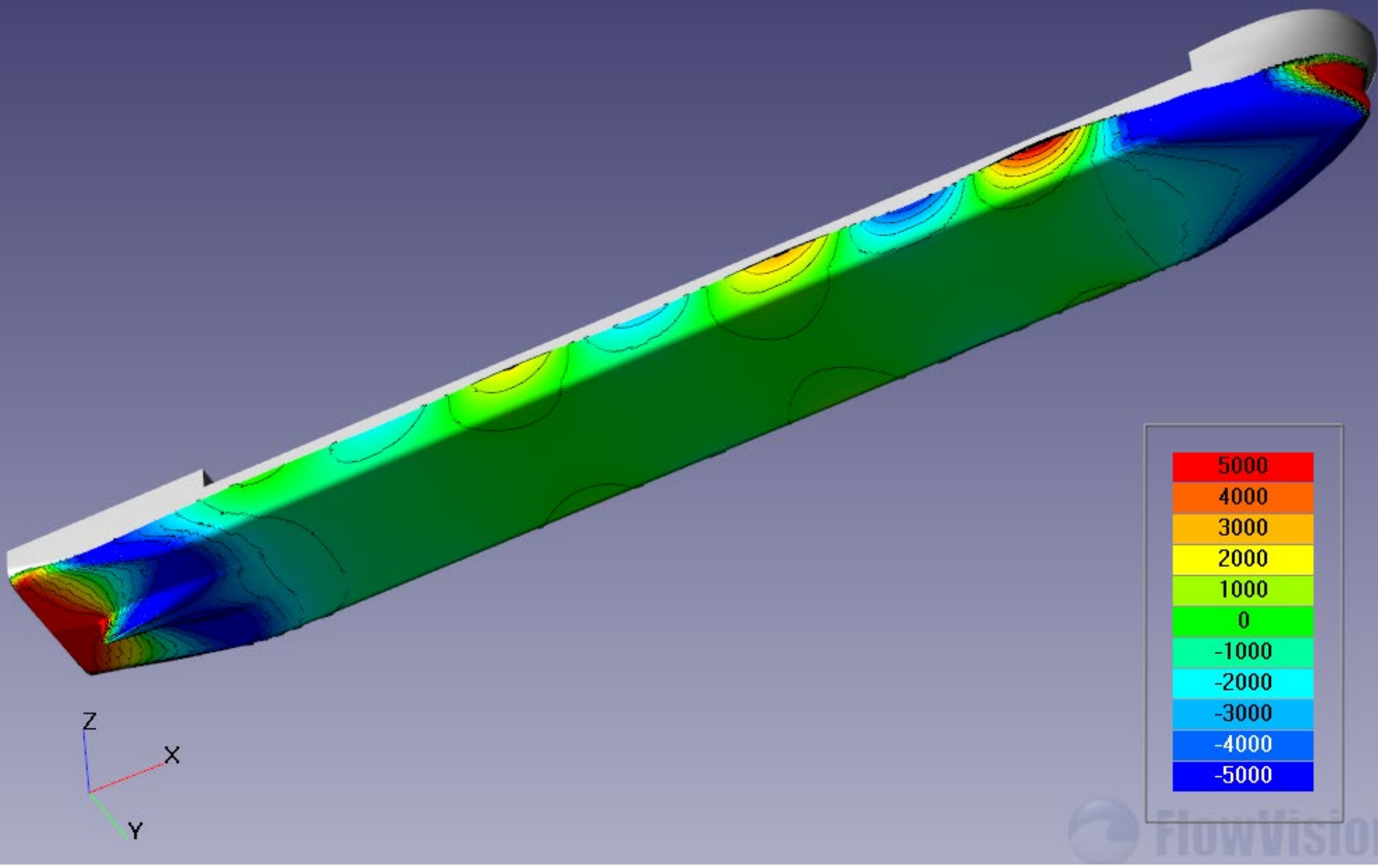
5.B - A practical approach

The technology components of HPCaaS that enable remote access to centralized resources in a multi-tenant way and their metered use are not unfamiliar to the HPC research and engineering community. However, as service-based delivery models take off, users have been mostly on the fence, observing and discussing the potential hurdles to its adoption. What is fairly certain is that we now have the technology ingredients to make it a reality. Let's start by defining what roles each stakeholder (industrial end-users, resource providers, software providers and high performance computing experts) has to play to make service-based HPC come together:

The industry end-user - A typical example is a small or medium size manufacturer in the process of designing, prototyping and developing its next-generation product. These users are prime candidates for HPC-as-a-Service when in-house computation on workstations becomes too lengthy and acquiring additional computing power in the form of an HPC server is too cumbersome or not in line with IT budgets. Plus, HPC is not likely to be the core expertise of this group.

The application software provider - This includes software owners of all shapes and sizes, including ISVs, public domain software organizations and individual developers. The UberCloud Experiment usually prefers rock-solid software, which has the potential to be used on a wider scale. For the purpose of this experiment, on-demand license usage is tracked in order to determine the feasibility of using the service model as a revenue stream.

The resource provider - This pertains to anyone who owns HPC resources (such as computers and storage) and is networked to the outside world in an open way. A classic HPC center would fall into this category as well as a standard datacenter used to handle batch jobs or a cluster-owning commercial entity that would be willing to provide compute cycles to run non-competitive workloads during periods of low CPU-utilization.



Team 46 - CAE simulation of water flow around a ship hull. The picture shows the pressure distribution on the hull surface.

The HPC experts - This group includes individuals and companies with HPC expertise, especially in the areas of cluster management and software porting. It also encompasses PhD-level domain specialists with in-depth application knowledge. In the Experiment, these experts acting as team leaders, work with end-users, computer centers and software providers to help the pieces of the puzzle fit together.

For example, suppose the user is in need of additional resources to increase the quality of a product design or to speed up a product design cycle - say for simulating sophisticated geometries or physics or for running copious amounts of simulations for a higher quality result. This suggests a certain software stack, domain expertise and even hardware configuration. The general idea is to look at the end-user's tasks and software and select the appropriate resources and expertise that match specific requirements.

Then, with modest guidance from the UberCloud Experiment team, the user, resource provider and HPC expert will implement and run the task and deliver the results back to the end-user. The hardware and software providers will measure resource usage; the HPC expert will summarize the steps of analysis and implementation; the end-user will evaluate the quality of the process and the results, in addition to the degree of user-friendliness this process provides. The experiment orchestrators will then analyze the feedback received. Finally, the team will get together, extract lessons learned and present further recommendations as input to their case study.

5.C - An experiment in progress

For the 125 teams from 66 countries (with over 1,100 active and passive organizations involved), the end-to-end process of taking applications to the cloud, performing the computations and bringing the resulting data back to the end-user has been partitioned into 23 individual steps which the teams closely follow on the Basecamp collaboration envi-

ronment. An UberCloud University ("TechTalk") has been created providing regular educational lectures for the community. And the one-stop UberCloud Exhibit [10] offers an HPC Services catalog where community members can exhibit their cloud related services or select the services that they want to use for their team experiment or for their daily work. Besides, many UberCloud HPC Experiment teams publish their results widely (for example, the article by Sam Zakrzewski and Wim Slagter from ANSYS entitled: "On Cloud Nine" [14]). Finally, at the November 2013 Supercomputing Conference in Denver, The UberCloud received the HPCwire Readers Choice Award for the best HPC cloud implementation [15].

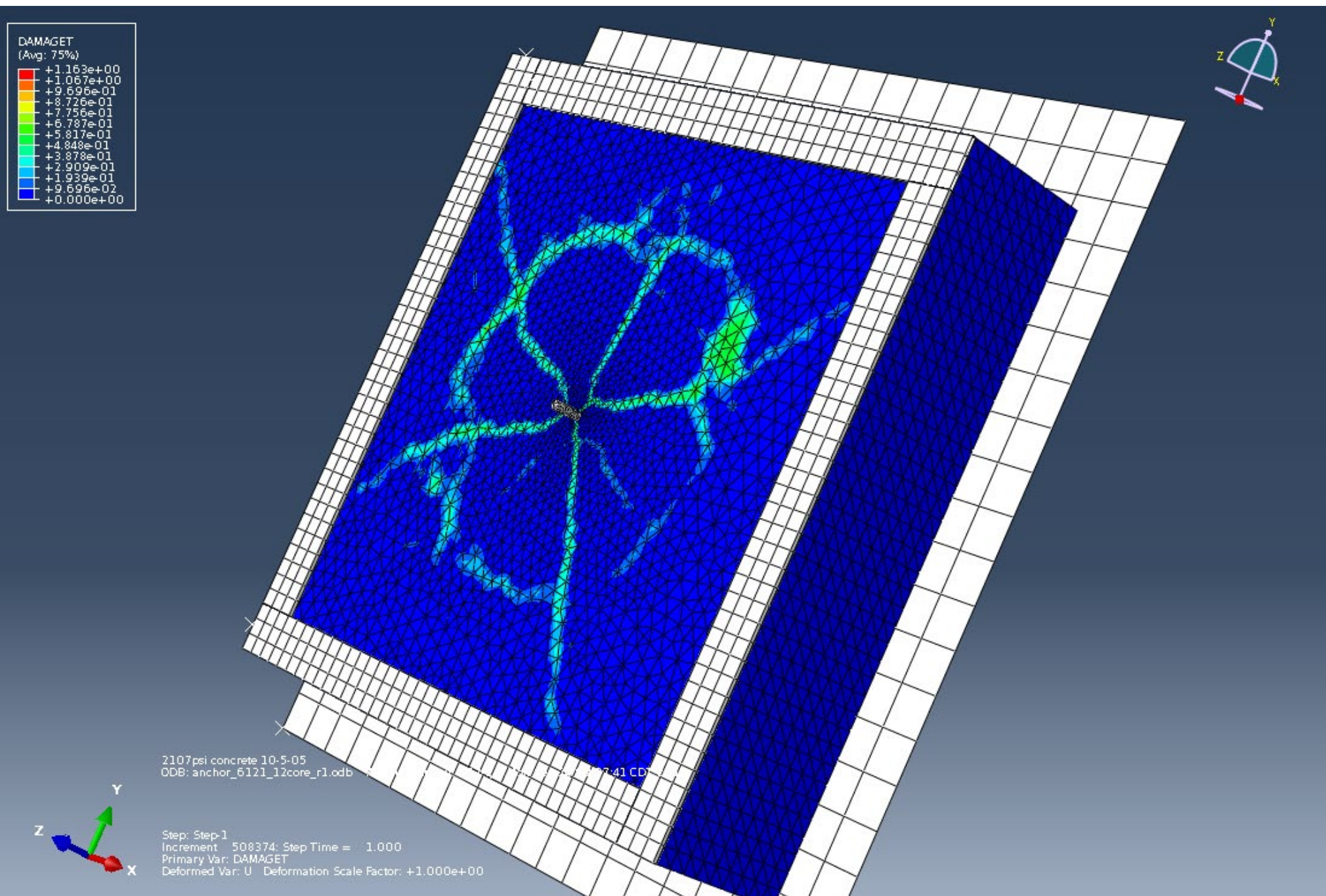
6 - HPC cloud Use Cases from the UberCloud HPC Experiment

As a glimpse into the wealth of practical use case results so far, we chose to present six out of 125 UberCloud Experiment results demonstrating the wide spectrum of CAE applications in the cloud (more in the Teams pages of the UberCloud's website).

Team 1: Heavy-duty structural analysis using HPC in the cloud.

This first team consisted of engineer Frank Ding from Simpson Strong-Tie, software provider Matt Dunbar with Abaqus software from SIMULIA, resource provider Steve Hebert from Nimbix and team expert Sharan Kalwani from (at the time of this experiment) Intel. The functions of this team range from solving anchorage tensile capacity and steel and wood connector load capacity to special moment frame cy-

Team 1 - Structural analysis model using HPC in the cloud. The picture shows concrete anchor bolt tension capacity simulation with 1.9 million degrees of freedom.

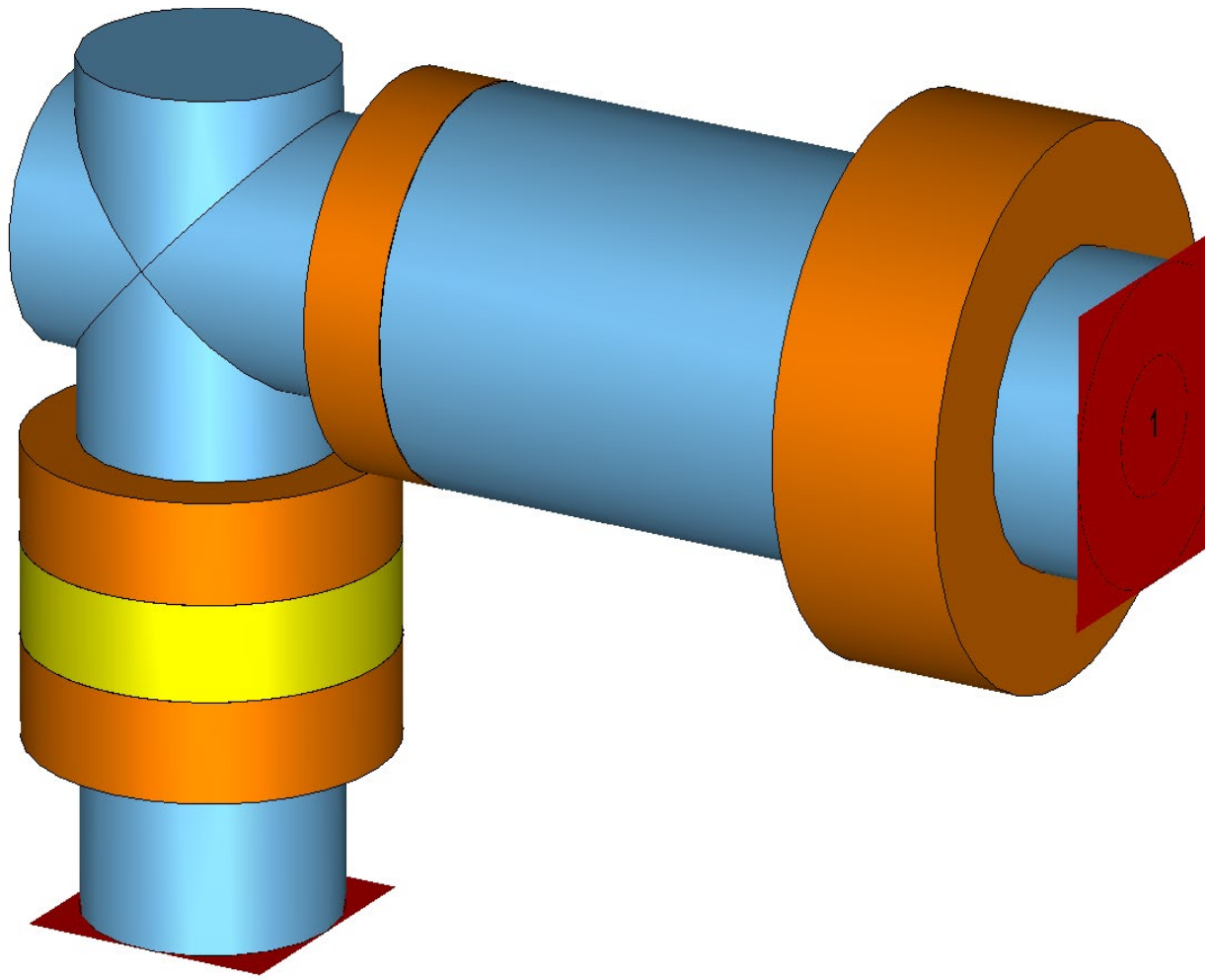


clic pushover analysis. The HPC cluster at Simpson Strong-Tie is modest (32 cores) so when emergencies arise, the need for cloud bursting is critical. Another challenge is the ability to handle sudden large data transfers, as well as the need to perform visualization to ensure that the design simulation is proceeding along the correct lines.

Team 2: Simulating new probe design for a medical device.

The end user's corporation is one of the world's leading analytical instrumentation companies. They use computer-aided engineering for virtual prototyping and design optimization for sensors and antenna systems used in medical imaging devices. Other participants in this team were software provider Felix Wolfheimer from Computer Simulation Technology, and Chris Dagdigian, CST and team expert from The BioTeam, Inc. The team used Amazon AWS Cloud resources.

From time to time, the end-user needs large compute capacity in order to simulate and refine potential product changes and improvements. The periodic nature of the computing requirements makes it difficult to justify the capital expenditure for complex assets that will likely end up sitting idle for long periods of time. To date, the company has invested in a modest amount of internal computing capacity sufficient to meet base requirements. Additional computing resources would allow the end user to greatly expand the sensitivity of current simulations and may enable new product and design initiatives previously written off as "untestable".



Team 2 - Simulation model of part of the probe to equip a new generation of medical device.

Hybrid cloud-bursting architecture permits local computing resources residing at the end-user site to be utilized along with Amazon cloud-based resources. The project explored the scaling limits of the Amazon EC2 instances and scaling runs designed to test computing task distribution via the Message Passing Interface (MPI). The use of MPI allows the leveraging of different EC2 instance type configurations.

The team also tested the use of the Amazon EC2 Spot Market in which cloud-based assets can be obtained from an auction-like marketplace offering significant cost savings over traditional on-demand hourly prices.

Team 8: Flash dryer simulation with hot gas Used to evaporate water from a solid

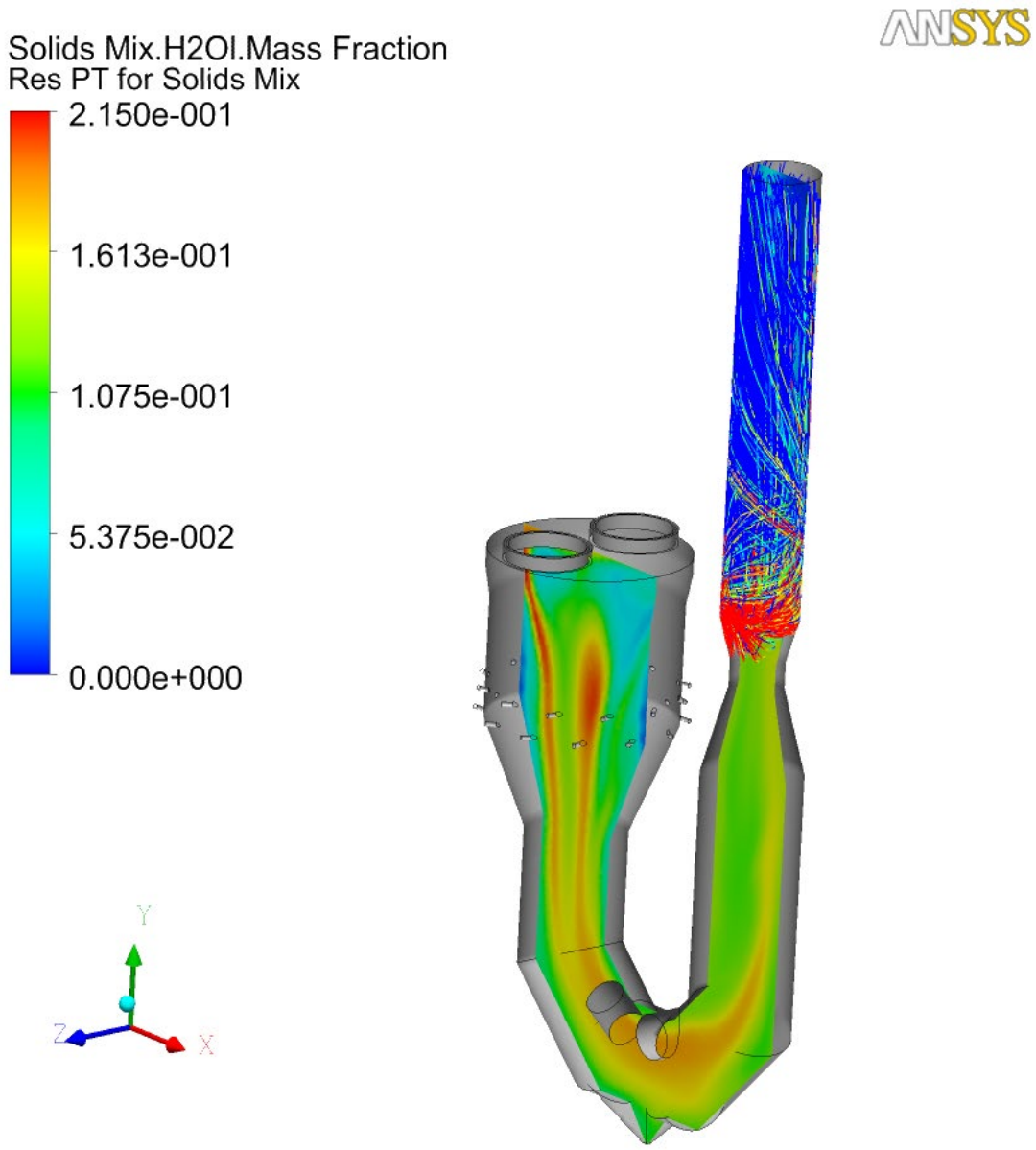
This team consisted of Sam Zakrzewski from FLSmidth, Wim Slagter from ANSYS as software provider, Marc Levrier from Serviware/Bull as resource provider and HPC expert Ingo Seipp from Science + Computing. In this project, Computational Fluid Dynamics (CFD) multiphase flow models were used to simulate a flash dryer using CFD tools that are part of the end-user's extensive CAE portfolio. On the in-house server, the flow model took about five days for a realistic particle-loading scenario. ANSYS CFX 14

was used as the solver. Simulations for this problem were using 1.4 million cells, five species and a time step of 1 millisecond for a total time of 2 seconds. A cloud solution allowed the end-user to run the models faster to increase the turnover of sensitivity analyses. It also allowed the end-user to focus on engineering aspects instead of using valuable time on IT and infrastructure problems.

Team 40: Simulation of Spatial Hearing

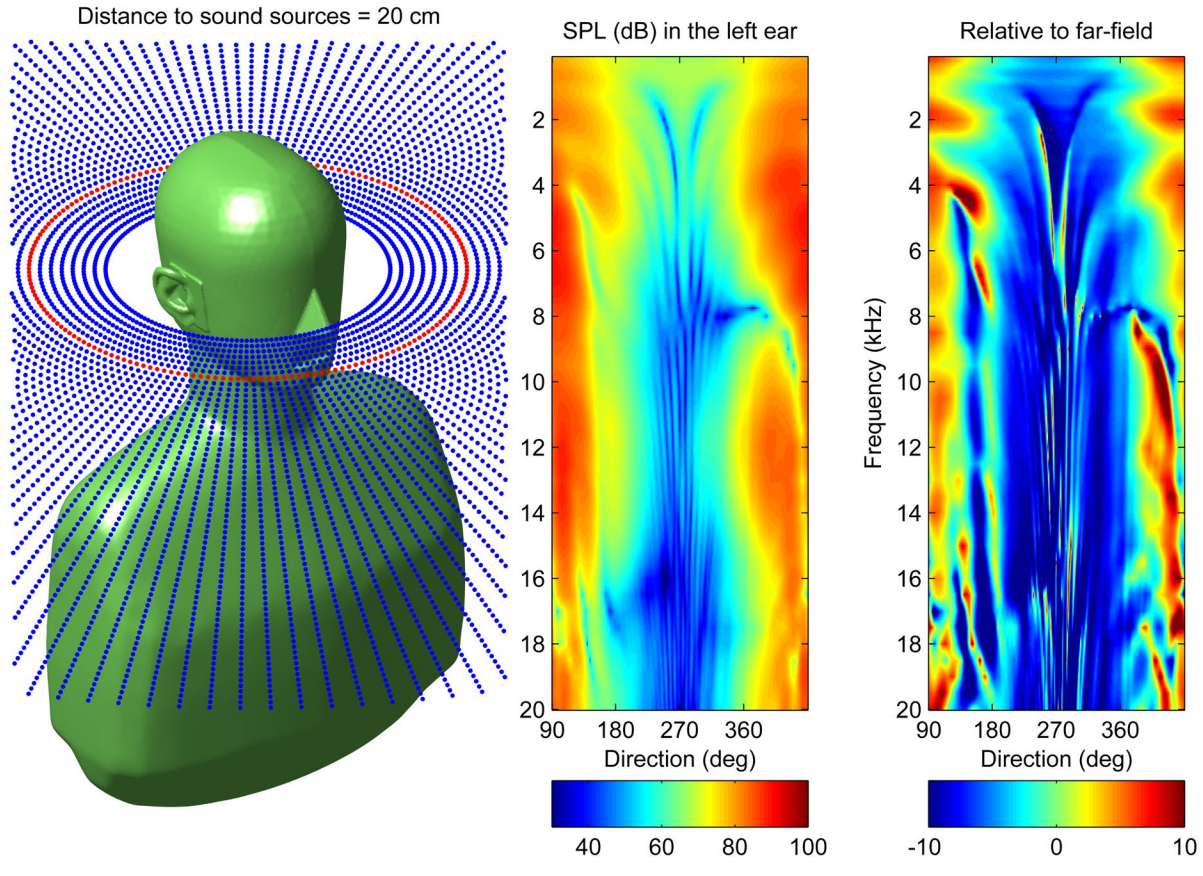
This team consisted of an anonymous engineer end-user from a manufacturer of consumer products, software providers Antti Vanne, Kimmo Tuppurainen and Tomi Huttunen from Kuava and HPC experts Ville Pulkki and Marko Hiipakka from Aalto University in Finland. The resource provider was Amazon AWS.

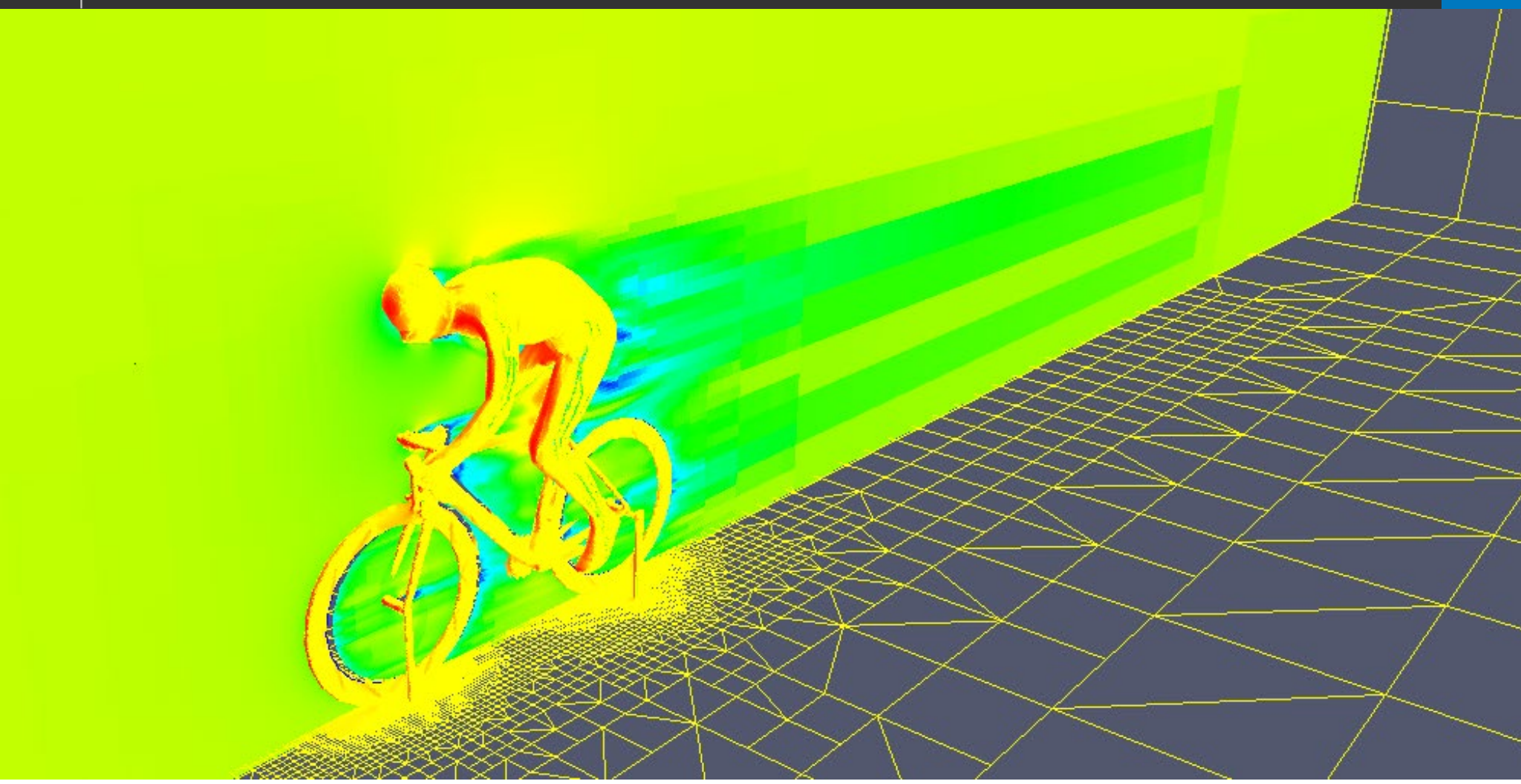
The human perception of sound is a personal experience. Spatial hearing (i.e. the capability to distinguish the direction of sound) depends on the individual shape of the torso, head and pinna (the so-called head-related transfer function, or HRTF). To produce directional sounds via headphones, one needs to use HRTF filters that "model" sound propagation in the vicinity of the ear. These filters can be generated using computer simulations, but to date, the computational challenges of simulating HRTFs have been enormous. This project investigated the fast generation of HRTFs using simulations in the



(Above) Team 8 - Flash dryer model viewed with ANSYS CFD-Post.

(Below) Team 40 - Dots in the simulation model indicate all locations of monopole sound sources used in the simulations. Red dots are sound sources. The middle section shows the sound pressure level (SPL) in the left ear as a function of the sound direction and the frequency. On the right, the SPL relative to sound sources in the far-field is shown.





Team 58 - Simulating wind tunnel flow around bicycle and rider.

cloud. The simulation method relied on an extremely fast boundary element solver.

Team 58: Simulating wind tunnel flow around bicycle and rider

The team consisted of end-user Mio Suzuki from Trek Bicycle, software provider and HPC expert Mihai Pruna from CADNexus and resource provider Kevin Van Workum from Sabalcore Computing. CAPRI to OpenFOAM Connector and the Sabalcore HPC Computing Cloud infrastructure were used to analyze the airflow around bicycle design iterations from Trek Bicycle.

The goal was to establish a greater synergy among iterative CAD design, CFD analysis and HPC cloud. Automating iterative design changes in CAD models coupled with CFD significantly enhanced the engineers' productivity and enabled them to make better decisions. Using a cloud-based solution to meet the HPC requirements of computationally intensive applications decreased the turnaround time in iterative design scenarios and significantly reduced the overall cost of the design.

Team 118: Conjugate heat transfer for the design of jet engines in the cloud

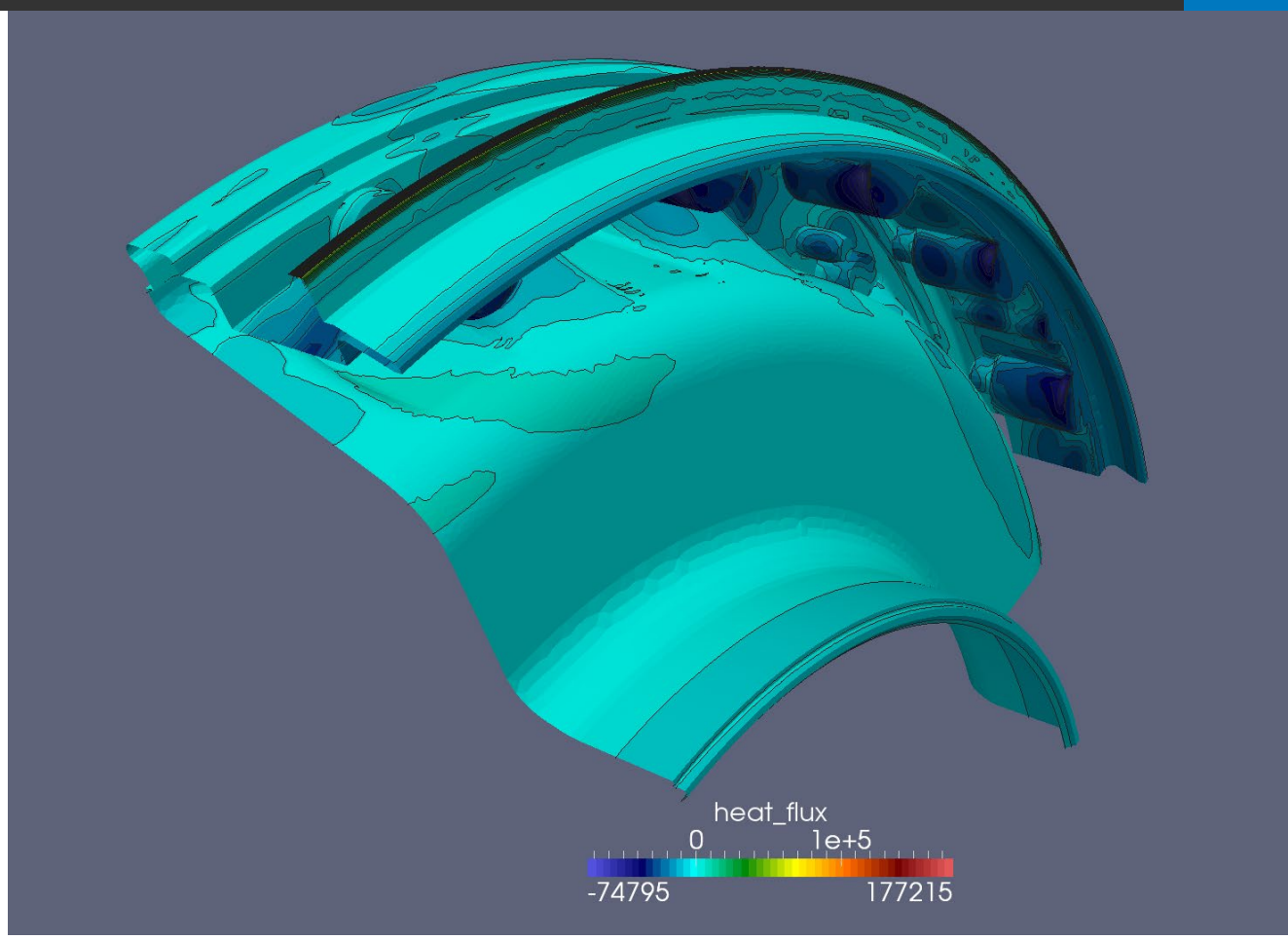
This team consisted of the end user, thermal analyst Hubert Dengg from Rolls-Royce Germany, the resource providers Thomas Gropp and Alexander Heine from CPU 24/7, software provider ANSYS, Inc. represented by Wim Slagter, and HPC/CAE expert Marius Swoboda from Rolls-Royce in Germany.

The aim of this HPC experiment was to link the commercial CFD code ANSYS Fluent with an in-house FE element code. This conjugate heat transfer process is very consuming in terms of computing power, especially when 3D-CFD models with more than 10 million cells are required. Consequently, it was thought that using cloud resources would have a beneficial effect regarding computing time.

Two main challenges had to be addressed: bringing software from two different providers with two different licensing models together, and getting the Fluent process to run on several machines when called from the FE software. After a few adjustments, the coupling procedure worked as expected.

From the end user´s point of view there were multiple advantages of using the external cluster resources, especially when the in-house computing resources were already at their

limit. Running bigger models in the cloud (in a shorter time) gave more precise insights into the physical behavior of the system. The end user also benefited from the HPC cloud provider's knowledge of how to setup a cluster, to run applications in parallel based on MPI, to create a host file, to handle the FlexNet licenses and to prepare everything needed for turn-key access to the cluster.



Team 118 - Contours of Heat Flux.

REFERENCES

[1] Council of Competitiveness, 'Make', 'Reflect', 'Reveal' and 'Compete' studies 2010.

[2] Randy Perry and Al Gillen. Demonstrating business value: Selling to your C-level executives. IDC White Paper, April 2007.

[3] NIST Cloud Definition

[4] Simon Aspinall. Security in the Cloud. Strategic News Service, 16-39, October 28, 2013.

[5] Steve Conway et al. IDC HPC End-User Special Study of Cloud Use in Technical Computing. June 2013.

[6] Jon Peddie. Workstations continue solid growth in Q3'13.

[7] Intersect360/NCMS study on Modeling and Simulation at 260 U.S. Manufacturers, July 2010.

[8] Wolfgang Gentzsch, Why Are CAE Software Vendors Moving to the HPC Cloud?

[9] Wolfgang Gentzsch and Burak Yenier. The Uber-Cloud Experiment. HPCwire, June 28, 2012.

[10] The UberCloud Services Exhibit.

[11] The UberCloud HPC Experiment: Compendium of Case Studies. Intel, June 25, 2013.

[12] Katherine Yelick, Susan Coghlan, Brent Draney, Richard Shane Canon. The Magellan Report on Cloud Computing for Science, Dec. 2011.

[13] Gartner Predicts Infrastructure Services Will Accelerate Cloud Computing Growth. Forbes Feb. 2013.

[14] Sam Zakrzewski and Wim Slagter. On Cloud Nine. Digital Manufacturing Report. April 22, 2013.

[15] The UberCloud Receives Top Honors in 2013 HPCwire Readers' Choice Award, November 27, 2013.

INTERNATIONAL
SUPERCOMPUTING CONFERENCE

ISC'14 THE HPC EVENT

June 22 – 26, 2014, Leipzig, Germany

Join the Global Supercomputing Community

www.isc-events.com/isc14

/verbatim

CATHERINE RIVIERE

PRESIDENT & CEO, GENCI
CHAIRWOMAN, PRACE COUNCIL

INTERVIEW BY STEPHANE BIHAN

On the agenda:

- The foundations of HPC in Europe
- Promoting simulation for SMEs' competitiveness
- Is Europe to be a global leader or follower in supercomputing?
- PRACE at Horizon 2020

Mrs. RIVIERE, as CEO of GENCI and Chairwoman of the PRACE Council, could you briefly present these two organizations and tell us specifically how they interact to develop high-performance computing in France and Europe?

GENCI (Grand Equipement National du Calcul Intensif, or National High Performance Computing Infrastructure) was created in 2007 by the French authorities following a 2005 report pointing out that France was lagging behind in providing computing capacity to our scientists, with the result that they were leaving to work abroad.

HPC is a strategic tool to maintain the competitiveness of a country. As such, the lack of resources presents a real

Twenty european countries have decided to join forces in order to implement a distributed computing infrastructure.

problem for academics and, in the long term, for industry. This is why it was necessary to put in place a genuine policy in this sector. The administration therefore decided to create a specific organization, GENCI, to place France at the highest level in Europe and internationally.

GENCI has three main missions: the first is to lead and support the national strategy for the provision of computing resources for scientific research. The second consists of representing France in any European governmental initiative, i.e. any initiative that requires national representation. Finally, GENCI's third mission is to promote simulation and HPC for research in both the academic and industrial sectors. GENCI is a non-trading company held by the Ministry of Research and Higher Education (at 49%), by CEA and CNRS at 20% each, by the CPU (Conference of University Presidents) (at 10%), and finally, by INRIA at 1%.

At the European level, PRACE (Partnership for Advanced Computing in Europe) did not yet exist when GENCI was founded, but its creation had already begun. Not only was it mentioned in the report on French policy on scientific computing, but the first Scientific Case for PRACE in 2005 had already proposed a roadmap for high performance computing at the European level. This document stated, *inter alia*, that at the time, no single country could compete with the United States or Japan. China was also on the list, but not at the level where it is today.

It was therefore essential that Europe mobilize. This was realized at both the national and European levels. France had a real problem of scientific competitiveness, and knowing that no member of the Union could compete with the major nations, we needed to unite. In 2008, the European Commission, which was of course anxious that Europe should play its role in high-performance computing, launched a preparatory phase project, PRACE, under the FP7 program, to develop an infrastructure for research at the European level. This initial preparatory phase project was followed in 2010 by implementation phases PRACE-1IP, 2IP, and 3IP, the current phase.

In budgetary terms, GENCI represents €30 million per year and PRACE €530 million over five years until 2015. These €530 million include €400 million from the four machine hosting countries - France, Germany, Spain and Italy - on one hand and, on the other hand, €130 million - €60 million from the Commission and €70 million from the 25 other representatives.

What is the total power of the computing infrastructure available to the European scientific community?

At the national level, 1.6 Pflops are distributed among three computing centers: the TGCC center at the CEA, the CINES center in Montpellier, and the IDRIS center at Orsay. Since 2008, GENCI has updated the equipment at these three centers to reach the 1.6 Pflops in question: 280 Tflops at CINES, the first GENCI installation in 2008; more recently, IDRIS with approximately 1 Pflops; and finally, the CURIE machine, which accounts for 2 Pflops. However, 80% of the latter are used for the French commitment to PRACE, the remaining 20% being counted as national computing resources.

MareNostrum in Barcelone (BSC).

At the European level, four countries provide 15 Pflops, corresponding to a total of €100 million per year for each of them. France provides the CURIE machine (CEA); Spain provides the MareNostrum machine (BSC) for 1 Pflops; Italy, the Fermi computer (CIneca) for 2 Pflops; and last, Germany provides access to three machines: Hermit in Stuttgart for 1 Pflops, JUQUEEN in Jülich for 6 Pflops and SuperMUC in Garching for 3.2 Pflops.

How is access to these resources organized?

Access to these resources is open to scientists and industry users based on the sole criterion of scientific excellence assessed by a Committee, the Access Committee, consisting of renowned scientists. PRACE offers three types of access: the first, called "preparatory" access, is intended to conduct tests for the second type of access, called “classic,” which extends over one year on the basis of computing hours granted. Finally, the last type of access is called "multiyear" and concerns large projects on a European scale.

Can you describe the PRACE infrastructure selection and construction process for us?

At GENCI, we issue traditional competitive calls for tender based on our users’ needs. As far as PRACE is concerned, the machines belong to the partners, and the partners are therefore free to choose the purchasing process for their machines. However, the goal is still to develop a European infrastructure consisting of different machines in order to meet all of the (very different) users’ needs.

So PRACE doesn’t finance the purchase of machines itself?

No. France, Spain, Italy and Germany, the four countries who were the first to bet on European computing, finance the purchase of these machines with their own funds.

Is what we are we talking about here only providing resources or is there also ongoing follow-up and support of scientific projects? Where appropriate, can you describe how user support is organized?

PRACE also meets its users’ needs with a service and support environment such as assistance in porting code. We set up six training centers whose purpose is to grow the computing ecosystem around PRACE supercomputers, together with the 25 other partner countries.

In order for Europe to remain competitive, PRACE has to become a sustainable infrastructure, which is not quite the case just yet.

HydrOcean, a start-up backed by GENCI through the HPC - SME initiative, received the prestigious IDC HPC Innovation Excellence Award for scientific excellence in fluid flows simulation. In concrete terms, what was this support?

This is a project that I am very attached to. HPC and digital simulation are essential tools to increase competitiveness, and so we needed to create a plan specifically for SMEs - the HPC-SME initiative. Promoting the use of simulation and HPC by industry falls squarely within GENCI's mission.

The objective is to help these SMEs ask themselves whether using simulation will increase their competitiveness. How can we help them to answer that question? With a pragmatic approach whose sole aim is to meet each SME's needs – not to sell them an innovative method. This implies listening and finding the most suitable solution for them, knowing that the final decision to use the resources is theirs. This initiative, out of the combined efforts of the GENCI and INRIA, was immediately supported by BPI France, which allowed us to launch the program as early as 2010.

So specifically, it consists of both financial and technical support?

You know, simply giving an SME the opportunity to compute on a machine is of no great interest, especially if it does not know a priori what that means. On the other hand, a proactive approach to the question allows us to ask SMEs whether digital simulation can help them, if so, how, and only then to show them how. It is a proof of concept, a real demonstration, with custom assistance. Thanks to their sound understanding of SMEs, competitiveness clusters, particularly those who include digital simulation in their approach, have naturally been involved.

The next step was to accompany SMEs with experts who are familiar both with the world of digital simulation and the business world, i.e. the EPICs such as IFP Energies Nouvelles and ONERA, who supported us right away, and the CNRS.

Overall, one must understand this strategy as a combination of competitiveness clusters, which disseminate and reach out at the regional level, and institutions that provide expertise in addition to GENCI's own network of experts. All of this exists in order to address all of the problems facing SMEs - most of them being unfamiliar with high performance computing.

However, the icing on the cake was the launch in 2010, as part of Future Investments, of the Equipex Equip@meso project to make simulation and HPC a vector of scientific and technological development. Under GENCI's leadership, Equip@meso was to have coordinated investments on the order of ten "mesocenters" (i.e. tier-2 regional supercomputing centers). However, by becoming coordinator of equipment that did not fall within its scope of action, GENCI's role was in some sense limited to that of banker.

We therefore proposed to Equip@meso that we help them win the Equipex tender by participating in scientific leadership in the network, by disseminating regionally the best practices that we had acquired at the national and European levels and, above all, by asking them to expand the HPC - SME initiative locally. In this way we succeeded in developing this initiative by relying on regional mesocenters.

SuperMUC in Garching (LRZ).

Finally, increasing the competitiveness of our companies truly consists of spreading the good word about digital simulation and HPC among them in order to lead them there from a practical and technological perspective. SMEs have trusted this approach because it is open and GENCI has remained neutral: we open the door to possibilities, but in the end, they are still in control of their own decisions.

And so, in addition to HydrOcean mentioned earlier, could you give us another example of success resulting from this approach?

Since 2010, we have supported no fewer than 43 SMEs with some outstanding success stories like HydrOcean: thanks to parallelization and access to national computing resources, it increased its software performance 500%. By working five times faster, it can therefore potentially earn five times more contracts. This really is efficiency, not human effort. It should be noted that HydrOcean is the first French SME to have benefited from 13 million hours of computing on a German PRACE machine. Which, by the way, allowed it to win a major contract across the Rhine.

To give you another example, I would like to mention Nexio Simulation, a specialist in aeronautics electromagnetism based in Toulouse. Thanks to the support of HPC-SME and the CALMIP mesocenter, a partner of Equip@meso, it has just gotten two major contracts in Japan. I would like to emphasize that we are talking about custom support and not a transfer of R&D. If it had not received free access to major computing resources, it would probably have been unable to respond to the call for tenders in Japan.

The HPC-SME initiative has attracted considerable interest among many European projects. We have therefore included it in the PRACE-3IP implementation phase in order to deploy it at the Union level; it will lead to a certain number of synergies without which nothing ambitious will be possible.

With the ETP4HPC in Europe and the HPC and Simulation call for proposals in France, European and French authorities are finally realizing the strategic role of HPC in innovation and competitiveness. In your opinion, what has led to this awareness?

It comes simply from the comparison with the United States, which invests millions of dollars in this sector each year. But it also results from the vigorous mobilization of the scientific community in France and Europe to alert the public authorities that our competitiveness, our innovative energy, and our international expansion are all falling behind. It took this huge communication effort to raise our leaders' awareness of the strategic role of digital simulation. We should also emphasize that from now on, high performance computing will be indispensable in many sectors. This is the case in health care, for example, where legislation prohibiting animal testing is a real catalyst for the simulation of living creatures and the creation of new molecules.

In this regard, how well do you think the national and European programs complement one another? In what ways aren't they simply competing with one another?

In Europe, the commonly used term is "subsidiarity." What can't be achieved at the national level is achieved at the European level. For the record, the greatest allocation of hours in the world was made by PRACE on the German machine Hermit for an English team from the Met Office: no less than 144 million hours were granted to enable this team to gain three years' lead in the development of high resolution climate models. Clearly, even our American friends had not made such an allocation. Without PRACE, this project would not have been possible. Computing cycles are allocated to projects that are on a larger scale than those that we can undertake at the national level. PRACE has a Scientific Council, an independent body that judges projects on the basis of their scientific excellence, and that made the bet to allocate its computing hours to important projects simply because the request was exceptional. This is how Europe draws its partner countries upwards and onwards, by making scientific investments that are more easily achievable at the European level than at the national level.

Let's take a minute to focus on our governing bodies. What are the ambitions of high performance computing policies? And do you think the allocated resources are adequate with respect to these ambitions?

At the European level, Europe does have a genuine political will. This is even the first time that such a will has been displayed in the HPC area. The Commission wants to have a Europe that is competitive on the international stage, and we would hope that the Horizon2020 program will yield the results hoped for, even if, considering the €15 million allocated in the 2014-2015 Work Programme, the financial resources are not realistically adequate today. However, all is not settled, and it is important that the PRACE model become a sustainable European infrastructure, which is not quite the case just yet. The financial commitment for PRACE will run until 2015 - that is to say tomorrow. However, we are building PRACE's future today, and we will need a different form of financial support, because four countries alone will not be able to support access to HPC resources for all the countries of the Union. Other sources of funding will have to be found, whether they come from the European Commission or other partners, to support both the investment in the machines and their operating costs.

At the French level, when GENCI was founded in 2007, the initial budget was €27 million, and the plan was to double it over the next three years. Today it is still €30 million. We are of course fully aware of the difficulties related to the current economic context; they underline the absolute necessity to unite at the European level.

PRACE's role as a pillar of the development of HPC in Europe being now fully recognized, how is it positioned with respect to the ETP4HPC?

The policy of the European Commission concerning HPC is clear. It is a policy based on three pillars: a pillar of research infrastructure, an R&D pillar to build technologies

"PRACE will offer 50 Pflop/s machines by 2020, with complementary architectures to allow for new HPC uses...

and a third pillar concerned with applications. ETP4HPC is the technological pillar, PRACE the infrastructure pillar, and the Centers of excellence, still under construction, the applications pillar.

PRACE has just articulated its strategic vision through the year 2020. Could you present its main points and specific objectives for us?

As we just stated, PRACE has guaranteed funding until mid-2015. We have now defined the next phase with the aim of meeting the needs of users. Written by the Scientific Steering Committee on the basis of the Scientific Case that identifies these needs (among others), the PRACE Council agreed to aim for overall computing power on the order of 50 Pflops by 2020, with complementary architectures. We also want to develop different uses for the machines and to facilitate code porting, a practice that did not really exist in the previous PRACE model.

On the other hand, today the Council is working on the financial model for PRACE's future which, for its part, is still under discussion - and for which there should be greater funding. The objective is to have a clear vision by the end of the year.

Do you have a doubt about the availability of these funds?

Caution is needed, but it's now clear that PRACE is a success. However, based on the TCO of a 1 Pflops machine, the cost of 50 Pflops of computing power is not of the same order of magnitude. Given these facts, we must come up with a new a new financial model for PRACE.

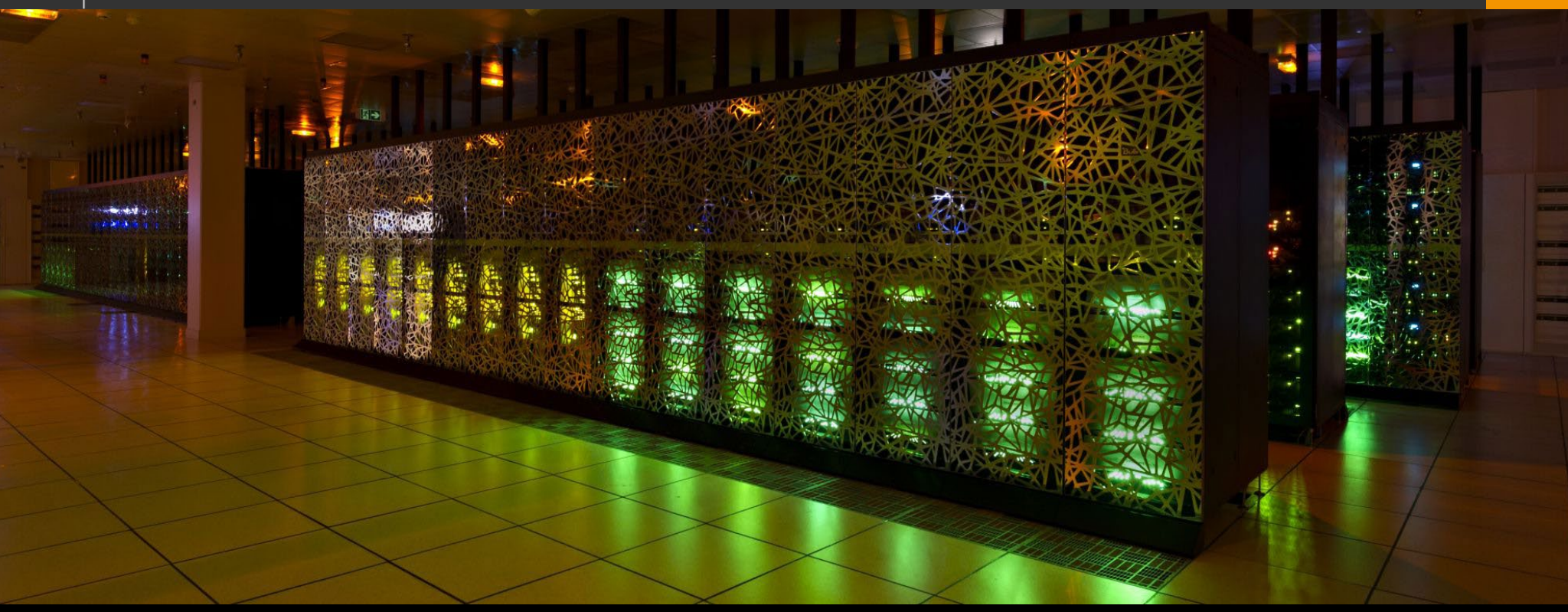
What are the challenges that Europe must now face to improve - or even maintain - its positioning in high performance computing at the international level?

The European high performance computing effort is a process that begins with ambitions, talent, and rather exceptional skills in the areas of use and software. Setting up such a process is time-consuming, and now the challenge is to make it thrive and develop. HPC is a global race in which historic players such as the United States and Japan are well placed, but also in which new emerging countries such as China, Russia and India are poised to catch up, or who have already caught up, with the leaders.

With respect to PRACE, thanks to the JUQUEEN machine, Europe has now reached 7th place in the Top 500 ranking. In terms of performance, 80% of the European public infrastructures are in PRACE. Europe is following the same trajectory in the increase in computing power, but at a level substantially below that of the United States. Europe must act as a "continent" and not follow a national approach, and it needs additional financing to reach that higher level.

In the race for the Exascale, Europe may appear "fragmented" compared to the United States or China, where computing centers are already focusing on achieving this objective by the end of the decade. Through its Horizon 2020 Program, does Europe aim to take the lead, or should it resign itself to follow its competitors?

To the question "Is Europe a follower?" I will answer "no," if only because the largest allocation in the world was



Curie at Bruyères le Châtel (TGCC - CEA).

made not in the United States or in China but in Europe. We are among the front-runners with many global advances achieved on European machines, such as the DEUS astrophysics project to model the structuring of the universe, carried out on the CURIE machine. From this perspective, Europe is not a follower.

Computing power is not really what's at stake here. What is at stake is the use that is made of it. Creating exaflop machines is a very interesting project, but how many teams will be able to use them? The challenge is really to develop uses, to train more people, and to give young people the urge to pursue technologies that people dream about.

PRACE must be seen as the actual demonstration of the will of Europe's leadership, or at least of its intent to place in this race to the Exascale, most likely with all the relative slowness inherent in our political and administrative organization, and with the fragmentation of our computing centers. Suffice it to say that the first discussions concerning PRACE, which was created in 2010, date back to 2004-2005. However, European high performance computing is up and running, and we are proud of this venture which has already led to some quite exceptional results. That is why the model begun and followed until now in PRACE must be perpetuated in PRACE 3, including a more integrated structure.

Along the same lines, in your opinion, what are Europe's strengths and weaknesses when it comes to HPC?

Europe has had very good scientists, know-how, and skills in the field of computing for quite a few years now. These are considerable strengths. However, funding is not at the same level as that allocated in the United States and China. That said, funding is not the essential factor: we must also build Europe to provide greater integration, efficiency, and responsiveness.

PRACE collaborates with XSEDE, its equivalent in some sense in the United States. Could you tell us more about this collaboration? How is it implemented? What are the objectives and expected results?

PRACE and XSEDE launched this initial collaboration focusing on development and tools testing in order to evaluate and enable interoperability between our two research infrastructures. A first call for projects was launched in late 2013, in response to which we received some very attractive proposals that we are currently evaluating.

Does PRACE intend to set up the same types of collaboration with China or other countries that are major international players?

It is of course necessary to talk with our Chinese and Japanese counterparts. For the time being, these talks are still quite informal. For example, we and the Japanese are currently discussing ways in which we

Fermi in Bologna (Cineca).

could work together and the scope of our collaboration. But remember that PRACE is still a relatively young organization that has yet to determine its processes and guidelines in this area. Being represented by twenty-five countries, four of which have greater voting rights, PRACE must obtain a consensus among its partners on the proper policy to pursue with China. Collaboration projects are underway, but we are moving forward carefully. In particular, we must make sure that we are all at the same level.

Isn't this relatively excessive caution with respect to the Orient regrettable?

China has a very clear line of action with a specific direction. They develop machines with their own processor technologies for applications they consider strategic. They have ambitions and the means to achieve them, both financial and intellectual. This excessive caution you are talking about is a political problem. What is the European policy towards China in HPC – and in other sectors? If HPC is considered as strategic in many industrial fields, it does necessarily call for a bit of caution...

In conclusion, two years after the installation of the CURIE machine, what will be the next "megasystem" placed at the disposal of the scientific community? What will its features be, and when will it be available?

At the national level, GENCI is renewing the computational capabilities of the CINES with a public tender that is underway. At the European level, the power of the SuperMUC machine will be doubled in early 2015 from 3.2 to 6.4 Pflops. The second German machine, Hermit, a CRAY XE6, will evolve into a CRAY XC30 4 Pflops system in late 2014 or early 2015. Regarding Spain's MareNostrum, its configuration will also evolve in 2015. As for Curie 2, whose launch we are targeting for 2017, it is a little early for all of the technical details to be finalized.

But you know, given the international context where supercomputers become obsolete so quickly, we have to do everything we can to promote high performance computing as a major component of Europe's scientific , economic and social strategy. This is the philosophy of our commitment, both at the national and continental levels. ■

Intelligent Rack PDUs

High Power and Intelligence precisely designed for HPC racks.

Choose from a broad portfolio of high power Intelligent PDUs:

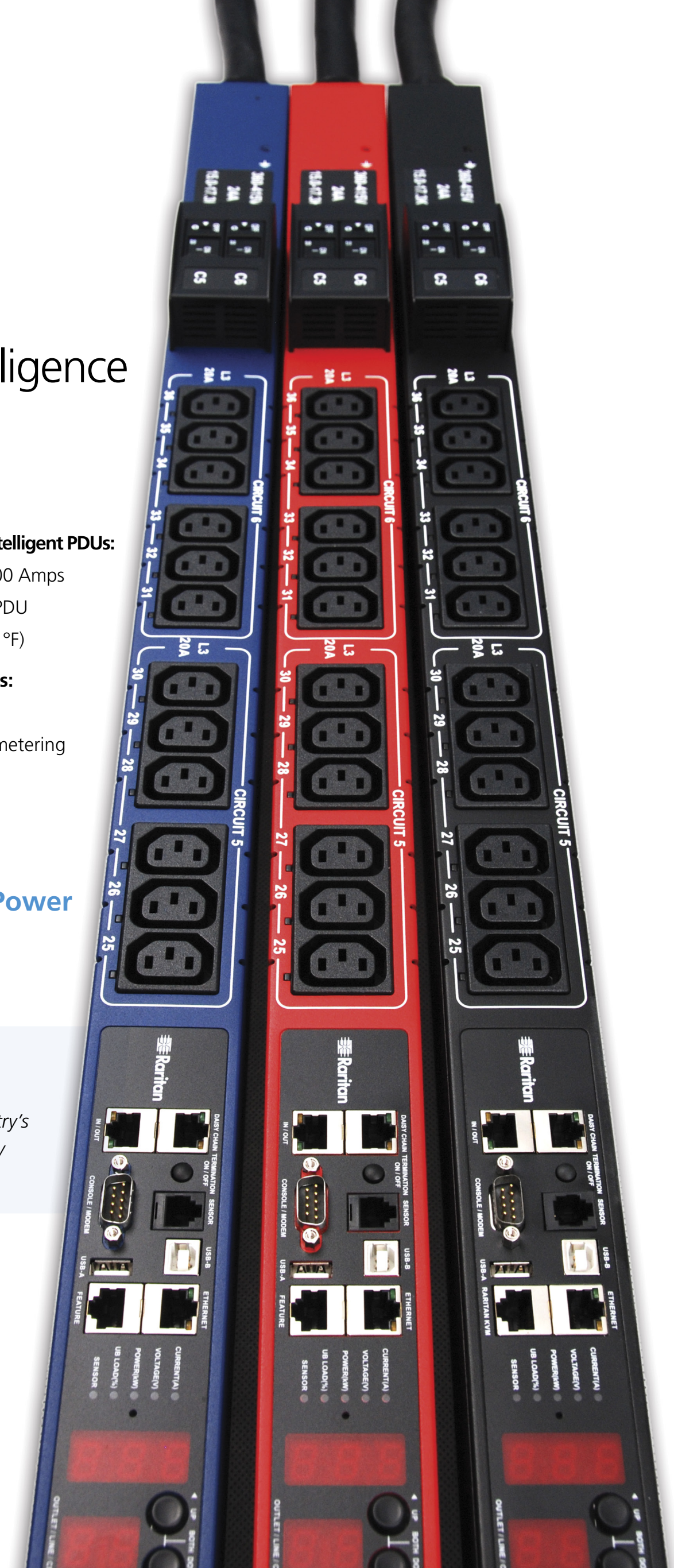
- High Power, capacities up to 55 kW and 100 Amps
- High Density, up to 54 outlets in a single PDU
- Highest ambient temperature (60 °C, 140 °F)

Leverage the industry's smartest capabilities:

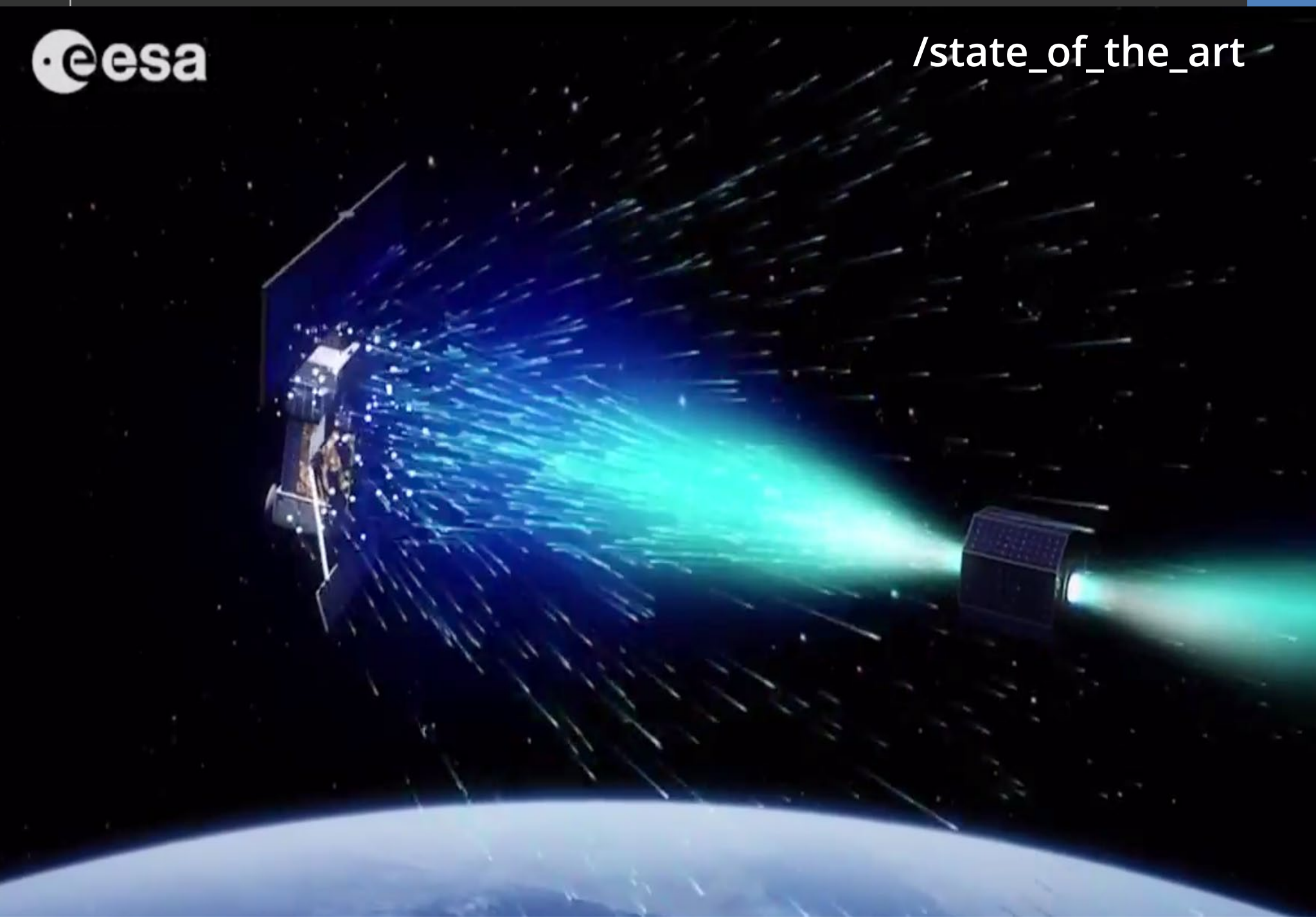
- Plug-and-play environmental sensors
- Accurate unit-level and outlet-level kWH metering
- Wi-Fi or wired networking
- Circuit breaker metering and monitoring
- Customizable to fit your HPC racks

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.™



SHIELDS TO MAXIMUM, MR SCOTT!

Researchers use TACC supercomputers to perform simulations of orbital debris impacts on spacecraft and fragment impacts on body armor. These simulations are evaluated at up to 10 kilometers per second to ensure that they accurately capture the dynamics of hypervelocity impacts and match real-world experiments conducted by NASA.

AARON DUBROW

We know it's out there, debris from 50 years of space exploration — aluminum, steel, nylon, even liquid sodium from Russian satellites — orbiting around the Earth and posing a danger to manned and unmanned spacecraft. According to NASA, there are more than 21,000 pieces of "space junk" roughly the size of a baseball (larger than 4 inches or 10 centimeters) in orbit, and about 500,000 pieces that are golf ball-sized (between 0.4 inch to 4 inches or 1 to 10 centimeters). Sure, space is big, but when a piece of space junk strikes a spacecraft, the collision occurs at a velocity of 3.1 to 9.3 mph (5 to 15 kilometers) per second—roughly ten times faster than a speeding bullet!

"*If a spacecraft is hit by orbital debris, it may damage the thermal protection system,*" warns Eric Fahrenthold, professor of mechanical engineering at The University of Texas at Austin, who studies impact dynamics both experimentally and through numerical simulations. "*Even if the impact is not on the main heat shield, it may still adversely affect the spacecraft. The thermal researchers take the results of impact research and assess the effect of a certain impact, crater depth and volume on the survivability of a spacecraft during reentry.*"

Only some of the collisions that may occur in low earth orbit can be reproduced in the laboratory. To determine the potential impact of fast-moving orbital debris on spacecraft — and to assist NASA in the design of shielding that can withstand hypervelocity impacts — Fahrenthold and his team developed a numerical algorithm that simulates the shock physics of orbital debris particles striking the layers of Kevlar, metal, and fiberglass that make up a space vehicle's outer defenses.

Supercomputers enable researchers to investigate physical phenomenon that cannot be duplicated in the laboratory, either because they are too large, small, dangerous — or in this case, too fast — to reproduce with current testing technology. Running hundreds of simulations on the Ranger, Lonestar and Stampede machines at the Texas Advanced Computing Center, Fahrenthold and his students have assisted NASA in the development of ballistic limit curves that predict whether a shield will be perforated when hit by a projectile of a given size and speed. NASA uses ballistic limit curves in the design and risk analysis of current and future spacecraft.

Results from some of his group's impact dynamics research were presented at the April 2013 meeting of the American Institute for Aeronautics and Astronautics (AIAA), and have recently been published in *Smart Materials* and *Structures* and *International Journal for Nu-*

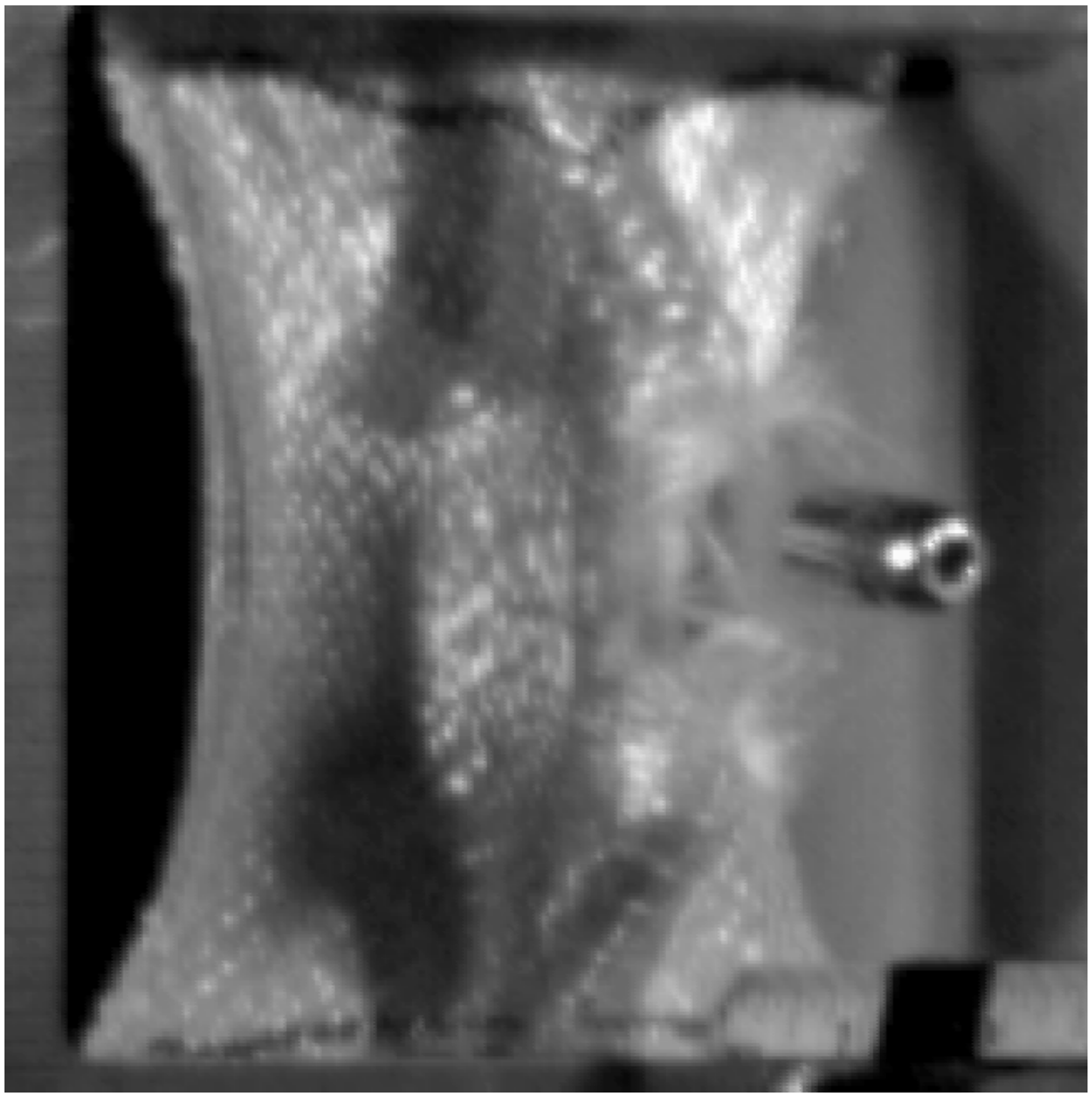
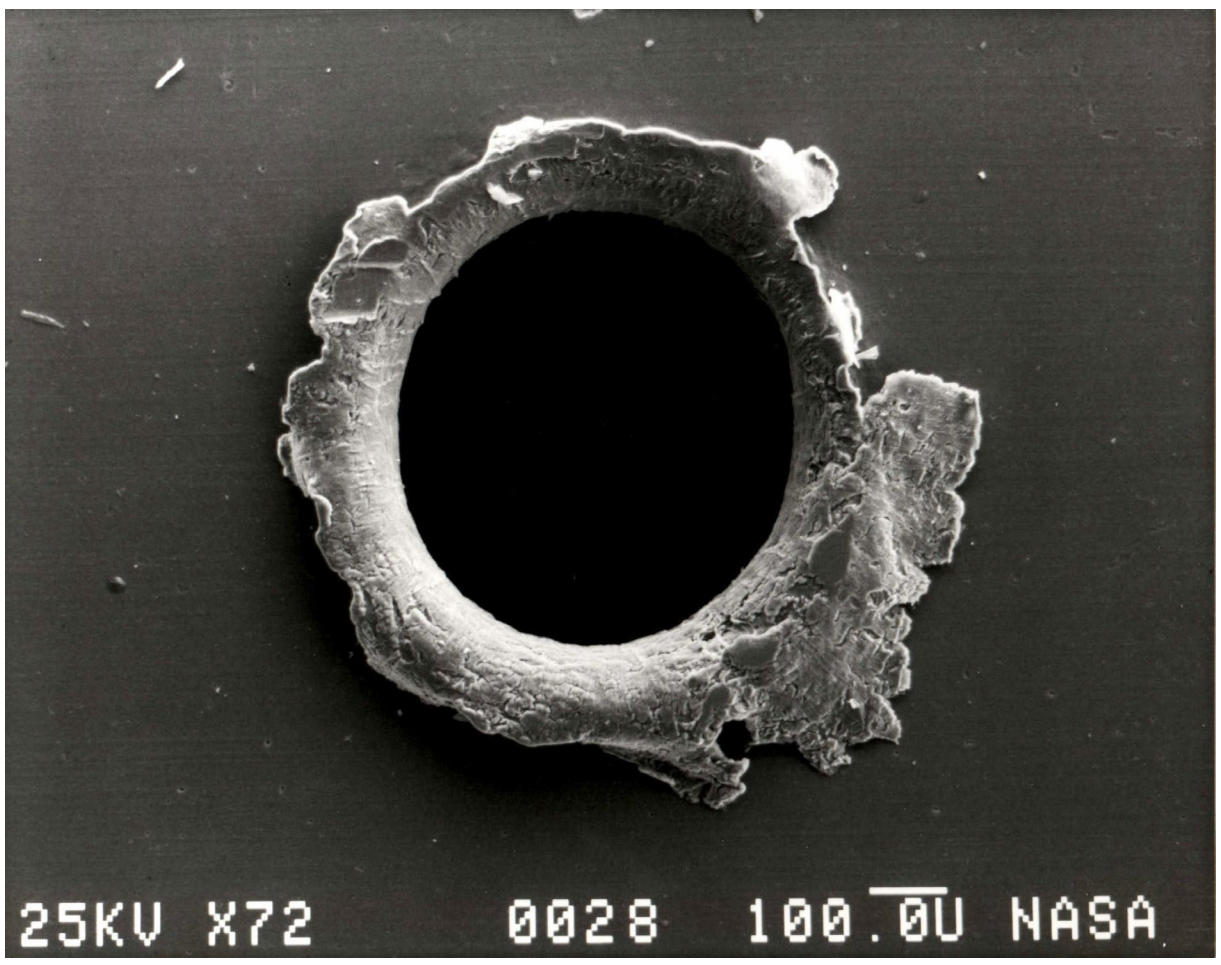


Fig. 1 - This high speed video frame depicts the perforation of a six-layer Kevlar target (4 inches/10.1 cm in width) by a 0.44 caliber copper projectile (experiment performed at Southwest Research Institute).

Fig. 2 - View of an orbital debris hole made in the panel of the SolarMax satellite. Credit: NASA, Orbital Debris Program Office



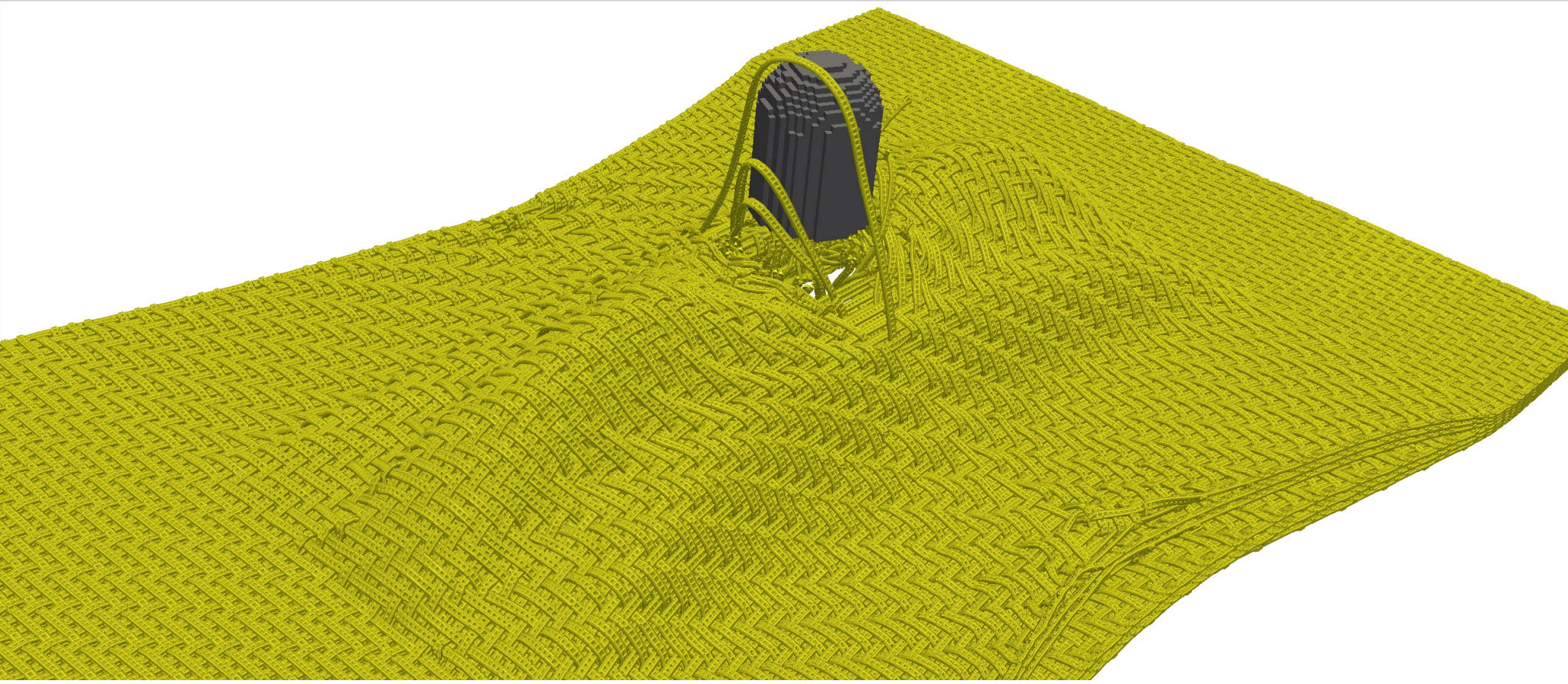


Fig. 3 - This simulation models the perforation of a six-layer harness satin weave Kevlar target (4 in/10.1 cm in width) by a 0.44 caliber copper projectile.

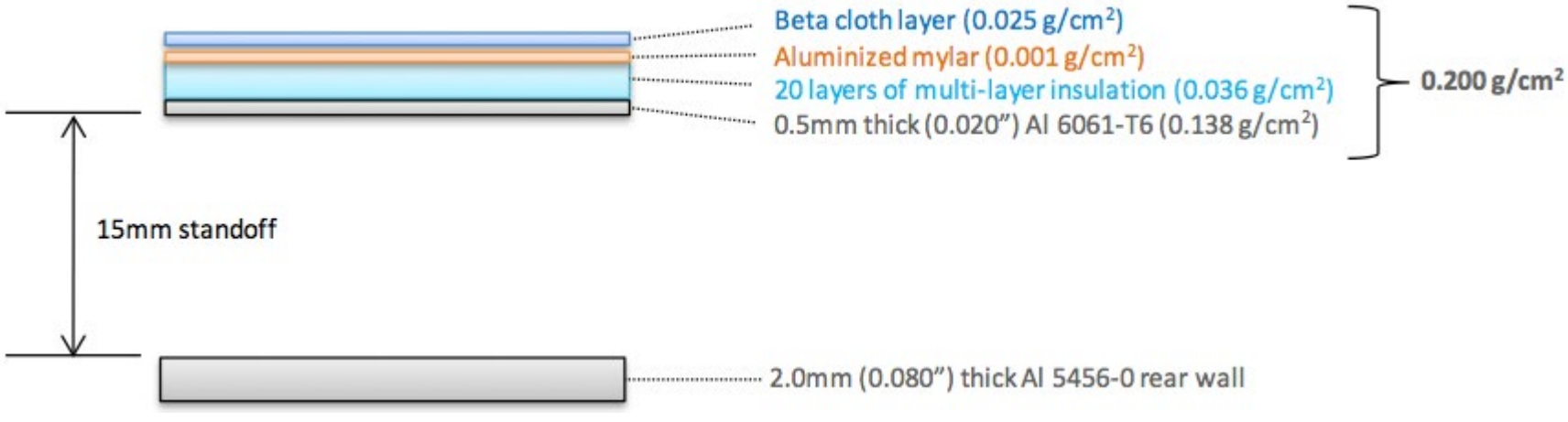
merical Methods in Engineering. In the paper presented at the AIAA conference, they showed in detail how different characteristics of a hypervelocity collision, such as the speed, impact angle and size of the debris, could affect the depth of the cavity produced in ceramic tile thermal protection systems.

The development of these models is not just a shot in the dark. Fahrenthold's simulations have been tested exhaustively against real-world experiments conducted by NASA, which uses light gas guns to launch centimeter-size projectiles at speeds up to 10 kilometers per second. The simulations are evaluated in this speed regime to ensure that they accurately capture the dynamics of hypervelocity impacts.

Validated simulation methods can then be used to estimate impact damage at velocities outside the experimental range, and also to investigate detailed physics that may be difficult to capture using flash x-ray images of experiments. The simulation framework that Fahrenthold and his team developed employs a hybrid modeling approach that captures both the fragmentation of the projectiles — their tendency to break into small shards that need to be caught — and the shock response of the target, which is subjected to severe thermal and mechanical loads.

"*We validate our method in the velocity regime where experiments can be performed, then we run simulations at higher velocities, to estimate what we think*

Fig. 4 - This figure (provided by NASA Johnson Space Center) shows an example of a spacecraft shielding system, which provides both thermal insulation and orbital debris protection in low earth orbit. The back (0.08 in/2 mm) aluminum layer is the wall of the spacecraft; in front (separated by a standoff) are shielding layers composed of aluminum, fiberglass, Teflon, and Mylar materials.



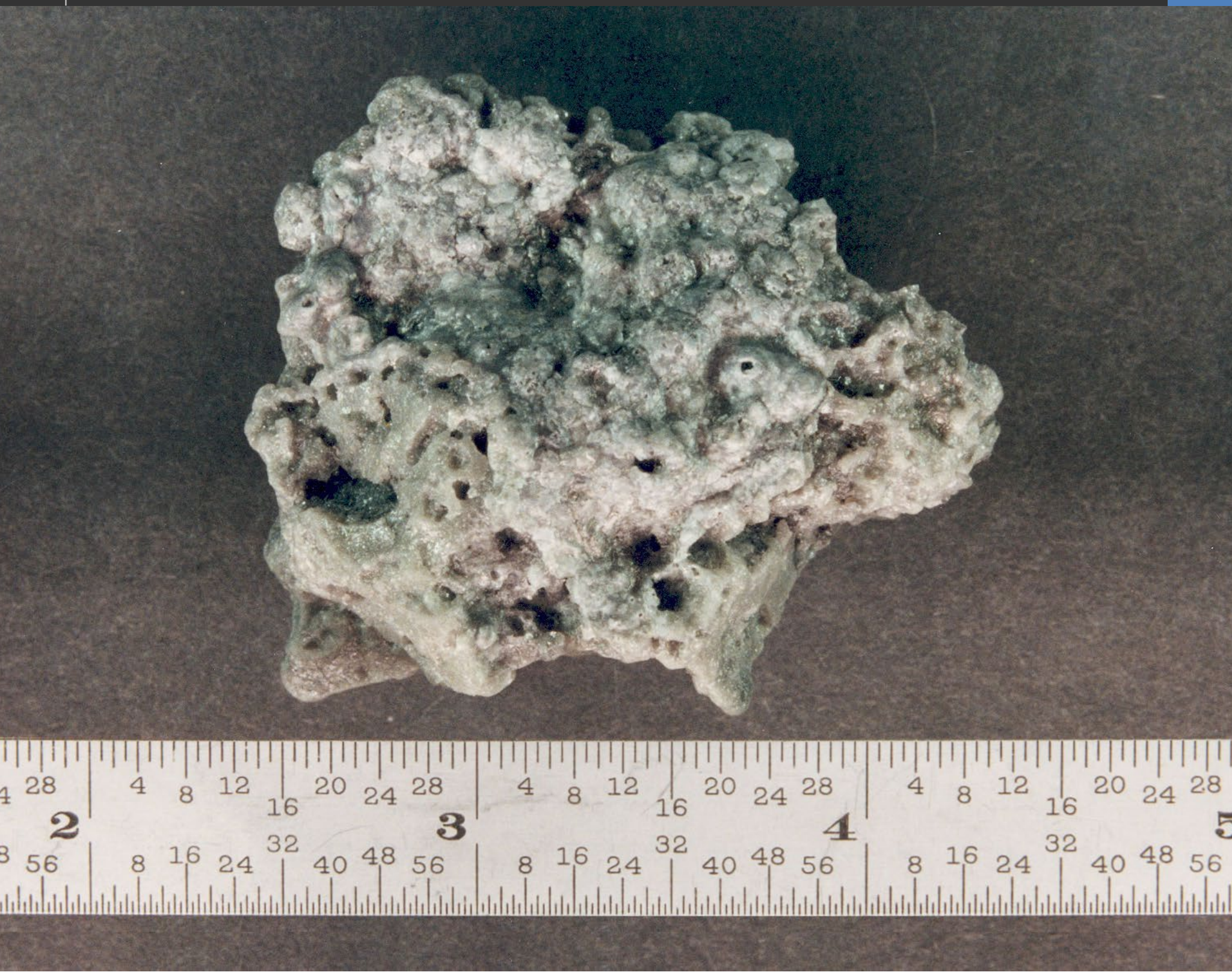


Fig. 5 - Solid rocket motor (SRM) slag. Aluminum oxide slag is a byproduct of SRMs. Orbital SRMs used to boost satellites into higher orbits are potentially a significant source of centimeter sized orbital debris. This piece was recovered from a test firing of a Shuttle solid rocket booster.

will happen at higher velocities," Fahrenthold explains. "*There are certain things you can do in simulation and certain things you can do in experiment. When they work together, that's a big advantage for the design engineer.*"

Back on land, Fahrenthold and graduate student Moss Shimek extended this hybrid method in order to study the impact of projectiles on body armor materials in research supported by the Office of Naval Research. The numerical technique originally developed to study impacts on spacecraft worked well for a completely different application at lower velocities, in part because some of the same materials used on spacecraft for orbital debris protection, such as Kevlar, are also used in body armor.

According to Fahrenthold, this method offers a fundamentally new way of simulating fabric impacts, which have been modeled with conventional finite element methods for more than 20 years. The model parameters used in the simulation, such as the material's strength, flexibility, and thermal properties, are provided by experimentalists. The supercomputer simulations then replicate the physics of projectile impact and yarn fracture, and capture the complex interaction of the multiple layers of a fabric protection system — some fragments getting caught in the mesh of yarns, others breaking through the layers and perforating the barrier. "*Using a hybrid tech-*

nique for fabric modeling works well," Fahrenthold says. "*When the fabric barrier is hit at very high velocities, as in spacecraft shielding, it's a shock-type impact and the thermal properties are important as well as the mechanical ones.*"

Moss Shimek's dissertation research added a new wrinkle to the fabric model by representing the various weaves used in the manufacture of Kevlar and ultra-high molecular weight polyethylene (another leading protective material) barriers, including harness-satin, basket, and twill weaves. Each weave type has advantages and disadvantages when used in body armor designed to protect military and police personnel. Layering the different weaves, many believe, can provide improved protection.

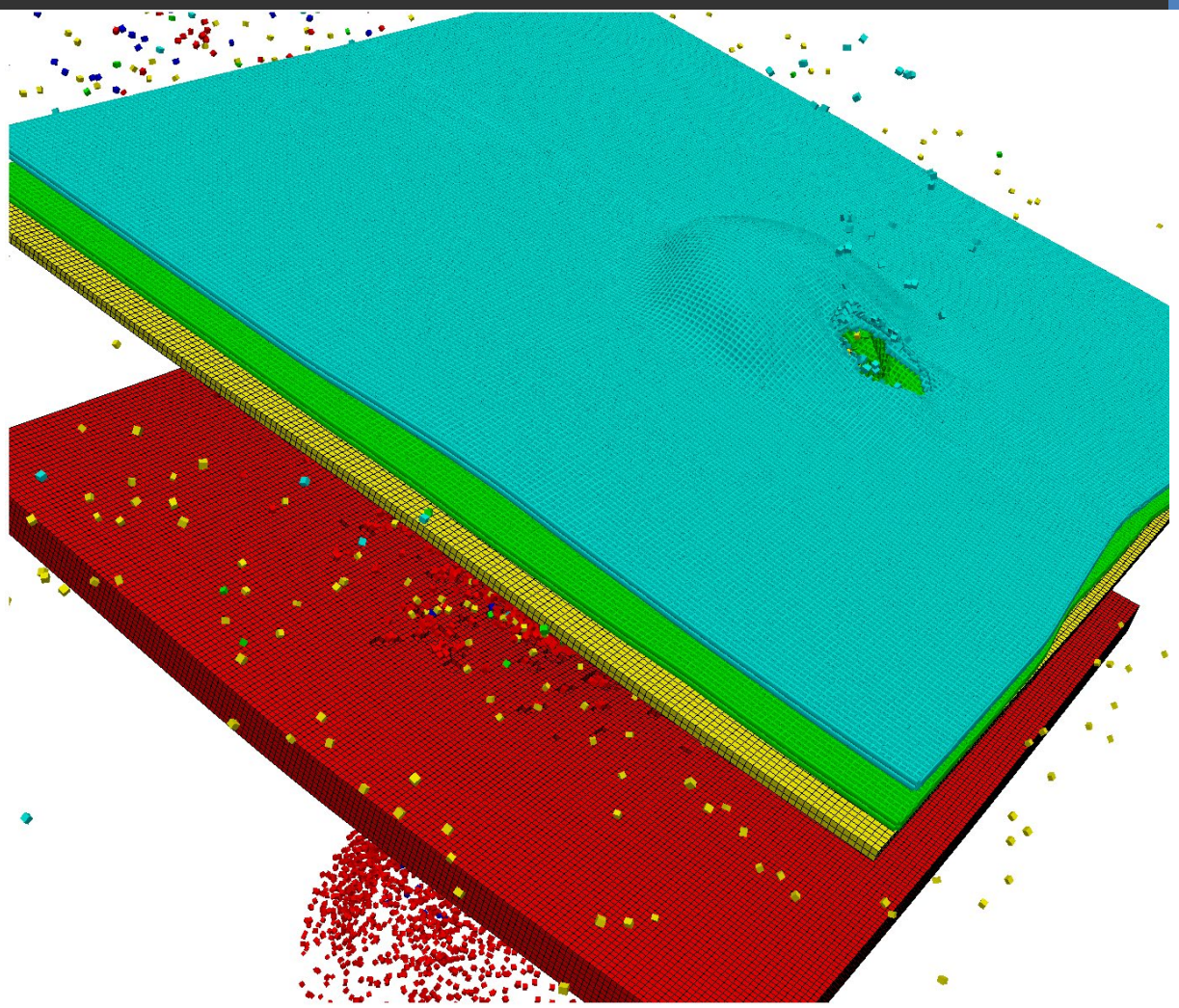


Fig. 6 - This figure shows a computational model of the spacecraft shielding and wall depicted above. The wall of the spacecraft is shown in red. The shielding is modeled as five layers: aluminum (in yellow), multilayer insulation (in green), and Beta-cloth plus aluminized Mylar (in blue). The model reproduces exactly the correct material areal densities, but the layer count is reduced, in order to reduce computational cost. Simulation results show the impact of a 0.94 in/0.24 cm diameter aluminum projectile on the spacecraft shield and wall. Impact velocity is 5.59 miles per second (9 km/s) and the impact obliquity is 30 degrees. The simulation indicates that this projectile will pass through the shielding system and perforate the wall of the spacecraft.

Fahrenthold and Shimek (currently a post-doctoral research associate at Los Alamos National Laboratory) explored the performance of various weave types using both experiments and simulations. In the November 2012 issue of the AIAA Journal, they showed that in some cases the weave type of the fabric material can greatly influence fabric barrier performance. According to Shimek, "*Currently, body armor normally uses the plain weave, but research has shown that different weaves that are more flexible might be better, for example in extremity protection.*" Shimek and Fahrenthold used the same numerical method employed for the NASA simulations to model a series of experiments on layered Kevlar materials, showing that their simulation results were within 15 percent of the experimental outcomes.

"*Future body armor designs may vary the weave type through a Kevlar stack,*" Shimek said. "*Maybe one weave type is better at dealing with small fragments, while others perform better for larger fragments. Our results suggest that you can use simulation to assist the designer in developing a fragment barrier which can capitalize on those differences.*"

What can researchers learn about the layer-to-layer impact response of a fabric barrier through simulation? Can body armor be improved by varying the weave type of the many layers in a typical fabric barrier? Can simulation assist the design engineer in developing orbital debris shields that better protect spacecraft? The range of engineering questions is endless, and computer simulations can play an important role in the 'faster, better, cheaper' development of improved impact protection systems. "*We are trying to make fundamental improvements in numerical algorithms, and validate those algorithms against experiment,*" Fahrenthold concludes. "*This can provide improved tools for engineering design, and allow simulation-based research to contribute in areas where experiments are very difficult to do or very expensive.*" ■

1 & 2 juillet/*july* 2014

École Polytechnique
Palaiseau - France

Le rendez-vous international

HPC & SIMULATION

The International Meeting

Forum Ter@tec 2014

SIMULER POUR INNOVER

INNOVATION BY SIMULATION

Inscription/*Registration* www.teratec.eu

Platinum Sponsors

Gold Sponsors

Silver Sponsors

FROM THE FARM TO THE HOME: DEM IMPROVES THE WORLD

Discrete Element Modeling (DEM) using Lagrangian Multiphase models has been applied to a multitude of industrial applications. With recent advances in the simulation of discrete particles and computing power, it is now being applied to the off-road vehicle sector.

TITUS SGRO*

Powerful computing hardware has allowed automotive companies to expand upon their design simulations, reaching out from the "core" optimization areas of external aerodynamics and engine design to improve upon and perfect virtually every aspect of an automobile. One particular group of vehicles that has lagged behind traditionally but is now racing to catch up is off-road vehicles. This category includes, in addition to ATVs, massive industrial applications like farming equipment and construction vehicles as well as household items such as lawn mowers and snow blowers. Recent technological advances in the simulation world allow this machinery to be modeled using discrete elements to replicate small particles, bypassing approximations and guesswork and going straight to a high-fidelity examination of the true operating conditions of these vehicles - exactly like modern cars.

While fluid flow has been a centerpiece of the Computer Aided Engineering (CAE) and Computational Fluid Dynamics (CFD) for decades, being able to predict the motion of individual particles has been an elusive utopia. The automotive sector has been the most demanding of these kinds of applications, seeking to understand and improve the com-

* Technical Marketing Engineer
CD-adapco

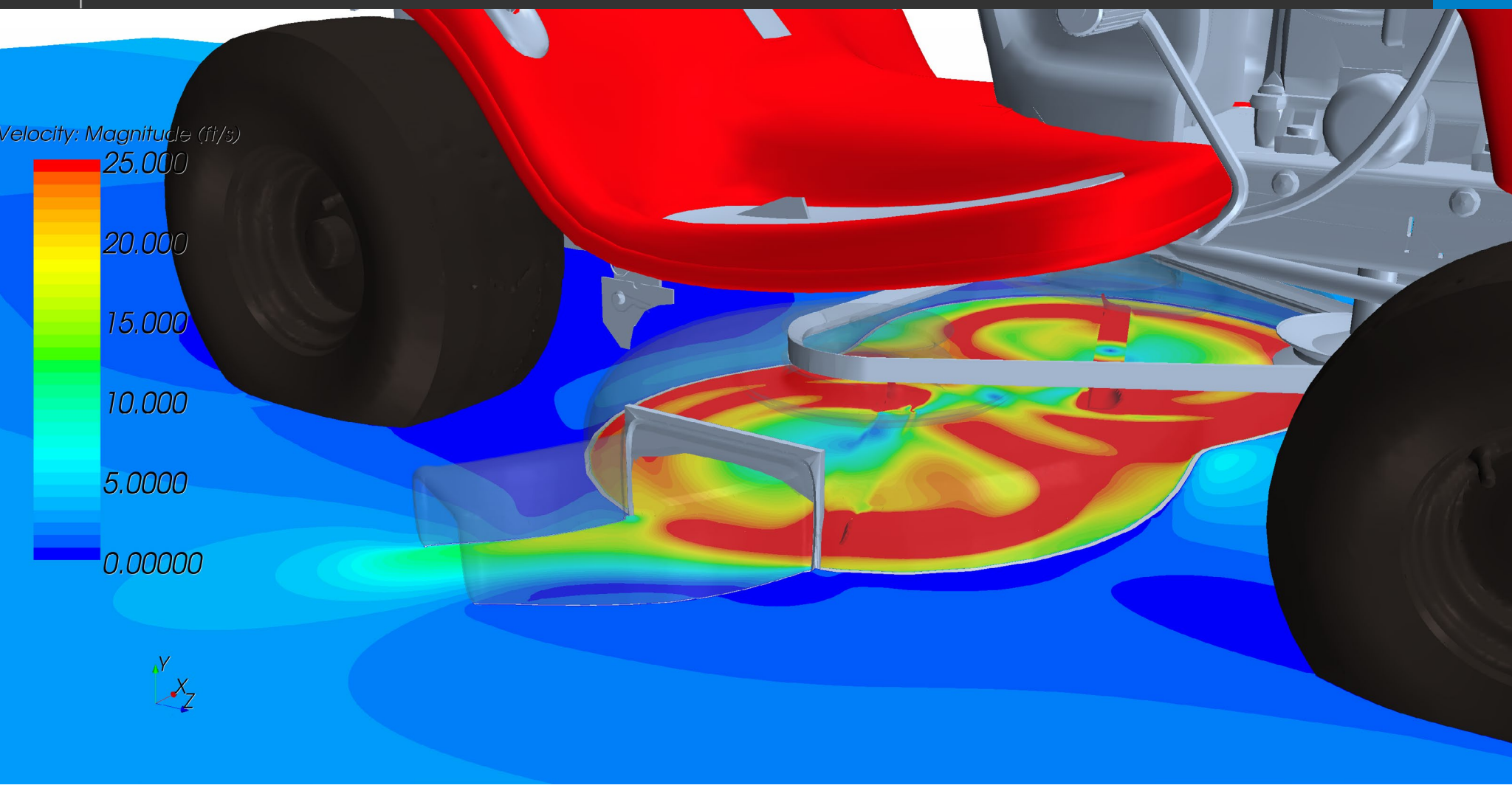


Fig. 1 - Simulation of airflow velocities within the mower deck.

plex processes going on within agricultural, construction and domestic vehicles. Agricultural and construction vehicles, (tractors, lawn mowers, front loaders, etc.) that are used to pick up and move enormous amounts of tiny particles, can be affected by airflow and motor mechanics, among many other physical mechanics. With recent advances in the simulation of discrete particles and computing power, engineers are finally able to use CAE to improve and optimize these vehicles by modeling each particle individually to create a much more faithful simulation of these complex machines.

Contaminant particles such as dust, as well as particulates that these objects are designed to work with, have previously been simulated in various methods as some form of fluid continuum. These particles include blades of grass (for a lawn mower), gravel and dirt (for construction equipment), and grain (for farming equipment). Simulating these particles as a fluid continuum creates imperfect results due to the simplification.

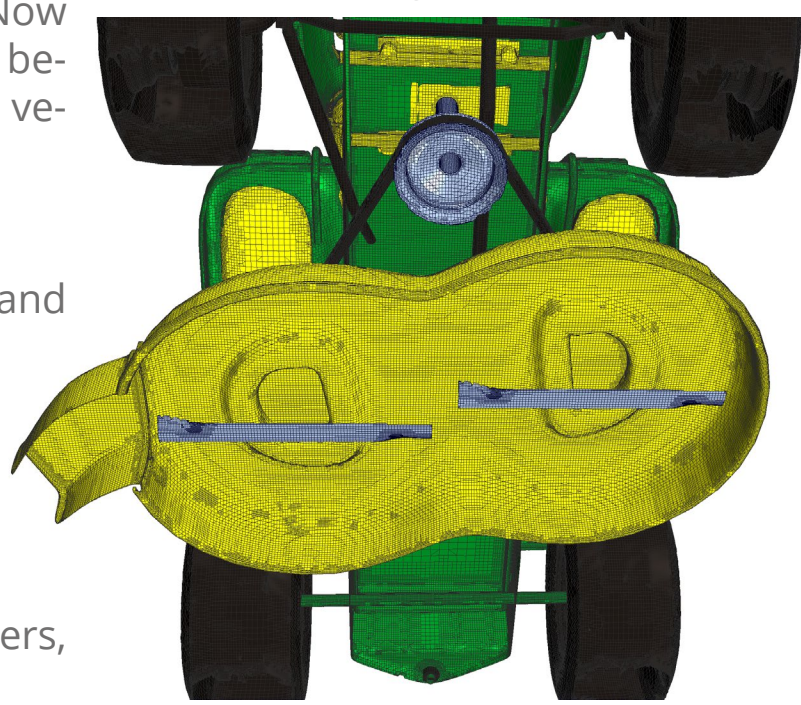
Applications of DEM

DEM using Lagrangian Multiphase has already been applied to a myriad of applications, from icing on airplane wings to mud and dirt simulation. Now this powerful technology is being applied to the off-road vehicle sector, including:

- Construction vehicles: dump trucks, excavators, and loaders
- Agricultural equipment: tractors and harvesters
- Domestic machinery: lawn mowers, snow blowers, and all-terrain vehicles
- Other areas of the Ground Transportation Industry: rain, mud, or snow modeling

A host of other industries have already begun using DEM, to simulate spray particles or atomizers – very common in the food industry, painting with aerosol cans or industrial applications for the automotive and

Fig. 2 - Underside view of volume mesh of lawn mower highlighting mowing deck and blades.



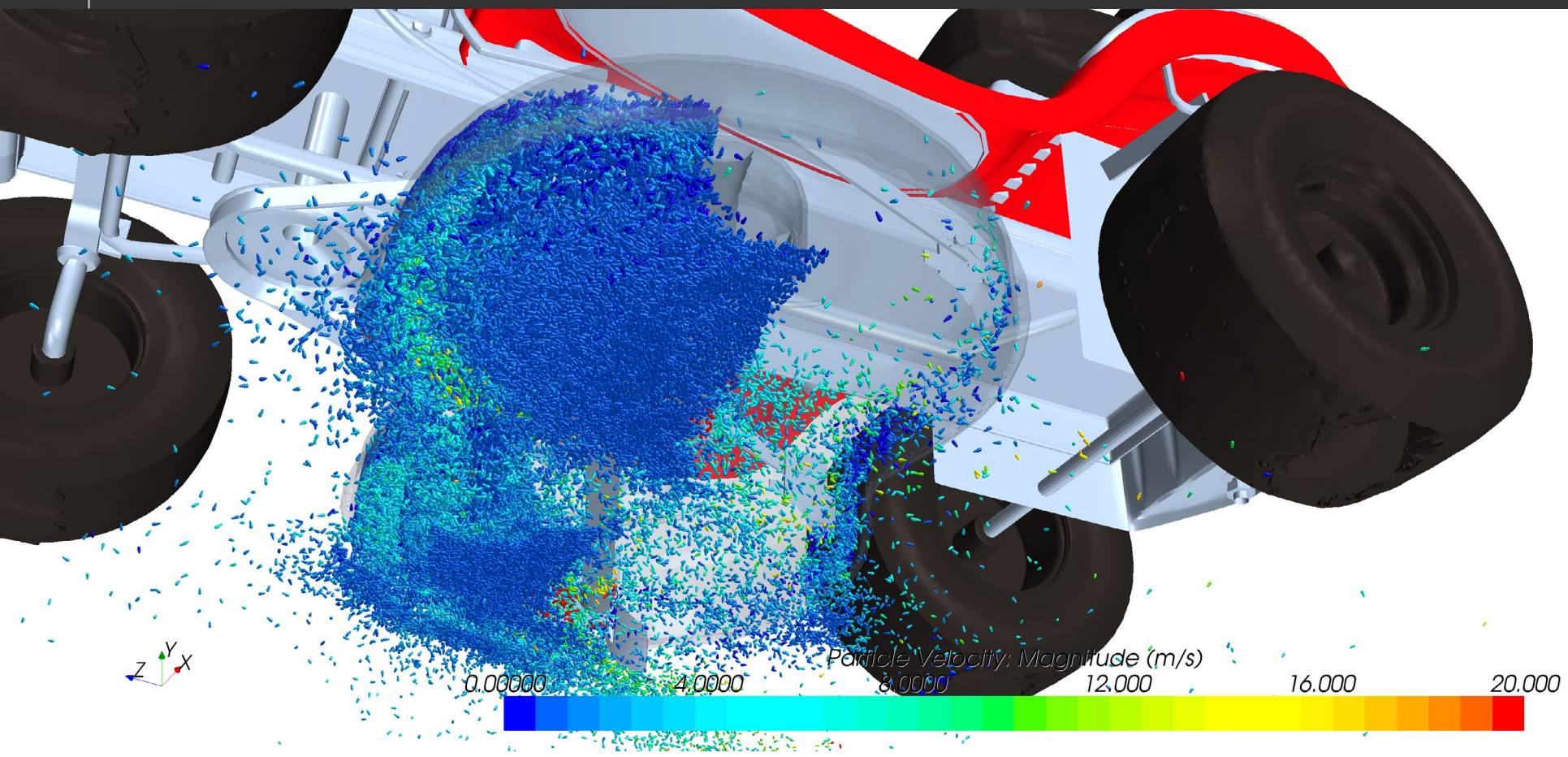


Fig. 3 - Underside view of grass particles in motion within the lawn mower deck.

aerospace industries – as well as the Life Sciences industry, investigating the dispersal of medication in a patient's blood from a pill or in the lungs from an inhaler.

Two blade tractor lawn mower

The following example uses a dual bladed tractor lawn mower, with the blades positioned under the driver, as seen in Fig. 2. A plate protects the driver from the grass being propelled at them. The blades were set in their own region with a separate mesh and spun using CD-adapco STAR-CCM+'s Rigid Body Motion (RBM), which rotates the entire region around a set axis. Because of the limitations of RBM, CD-adapco recommends that during each time-step of the rotation, the spinning region should be turned by 1 degree at most so that the rotating cells do not become completely out of line with the stationary cells. In the simulation pictured, the two meter long blades are spinning at 1,500 rpm, which translates into a time-step of 1.11×10^{-4}s. The simulation was set at 20 inner iterations per time-step to ensure convergence. The particle injector field consisted of a 15x15 rectangular grid of injectors and each injector was set to inject 100 particles per second. The simulation ran for a bit over 0.5s of simulation time, meaning that almost 13,000 particles were injected.

Lagrangian multiphase

The Lagrangian-Eulerian approach is based on "a statistical description of the dispersed phase in terms of a stochastic point process that is coupled with an Eulerian statistical representation of the carrier fluid phase." In other words, the main fluid phase is solved as a continuum using the time-averaged Navier-Stokes equations. The second, discretized phase, involves either a second fluid dispersed in relatively small pockets or small particulates of a solid that is mixed in with the liquid.

The Lagrangian Multiphase model makes use of DEM particles and simulates each particle individually as it is affected by the fluid domain and collisions with other particles and features within the simulation boundaries. This is the most faithful representation of particles possible within a fluid simulation, allowing for the examination of primary and secondary breakup models, as well as turbulent dispersion. However, as one can imagine, it also requires a lot of processor power, especially when the number of particles exceeds 105.

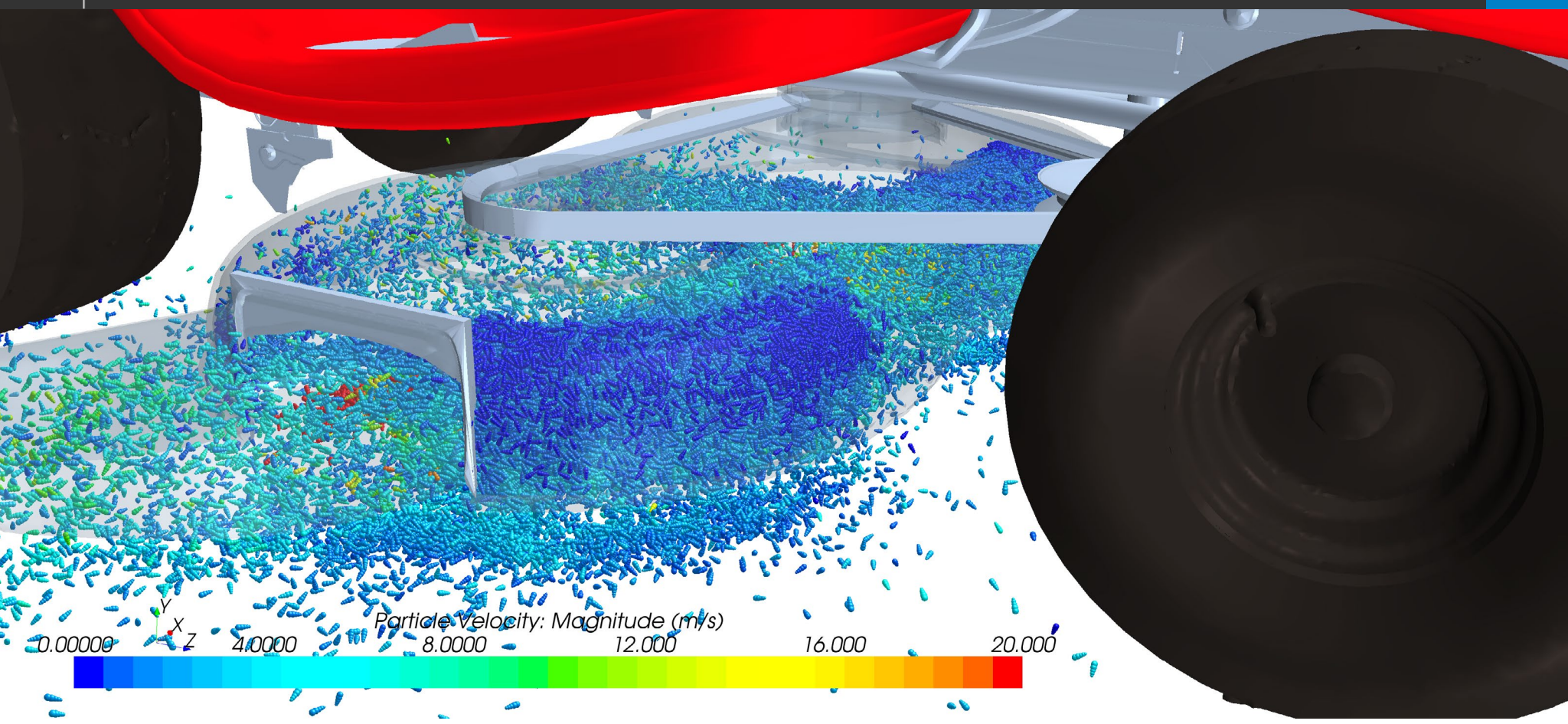


Fig. 4 - Grass particles in motion within the mower deck.

Certain assumptions and approximations were used within the simulation. The most important assumption was that the tractor lawn mower did not actually move within the simulation. The wheel rotation, translation and creation of a real grass field for the lawn mower to drive over being beyond the reasonable scope of the simulation, the movement was simulated by having the grass particles constantly injected into the cutting areas.

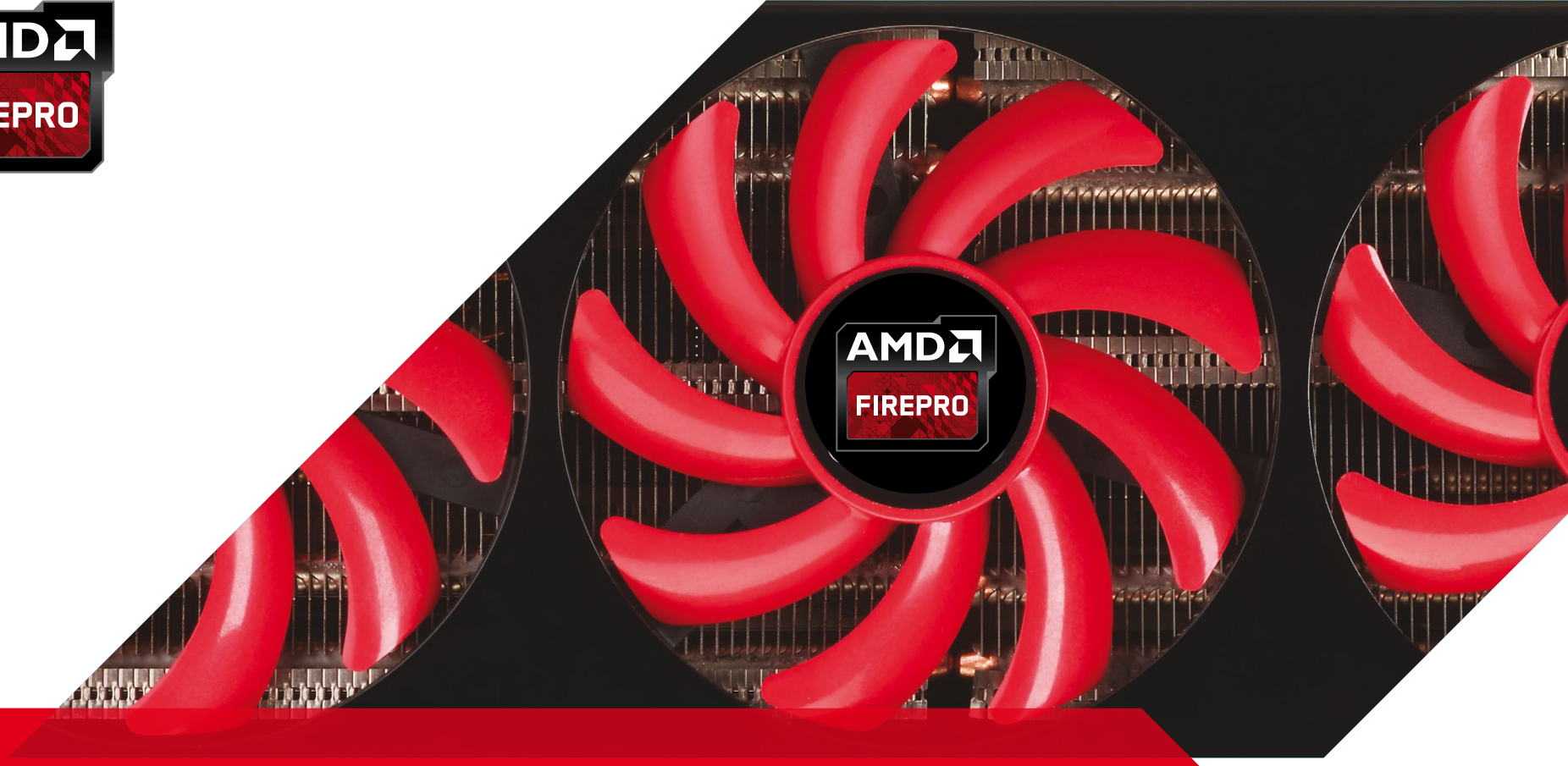
Since the full tractor body was not included in the simulation, it did not need to be meshed at all, allowing for finer refinement of the blades, protector plate, and air volume in the area of question. Another approximation was that the grass external to the plate was not physically present to impede the motion of the moving particles. Hence, an invisible barrier was setup around the perimeter of the plate to the ground to prevent the leakage of particles in areas besides the ejector. This barrier was set as a porous region, allowing the fluid (air) to flow freely but not allowing the particles to cross the barrier.

The third major approximation was the shape of the grass particles. Every grass particle was modeled as a composite of spherical particles in a 1 cm long conical shape. The composite particle was forbidden from being broken. To keep it simple and inexpensive, only one particle model was used. Additionally, the particles were injected into the region already separated from the ground; the lawn mower blades were not involved in separating them from the injector in any way and only provide air motion.

Because of the small time-step (caused by the 1 degree rotation mentioned above), a detailed picture of the motion of the particles was created. This enabled the visualization of areas where clumping occurred, as well as areas that were affected by internal vortices. Every particle was permitted to exit only through the ejector (in all of the figures, the ejector is positioned on the left-hand side). The particles were modeled with drag and their density was set to a value such that each particle would be primarily affected by the motion caused by the spinning blades.

As with the previous case study, the particles were colored with respect to their velocity magnitude. The benefit of this simulation was demonstrating that a complex physical simulation of a lawn mower can be accurately modeled, allowing for designers to see the movement of each individual blade of grass. This permits the designers to modify their design based on complexities too subtle to be seen with the naked eye, modifications which can significantly increase their efficiency by reducing fuel consumption and the time required to mow. ■

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

AN INTRODUCTION TO PERFORMANCE PROGRAMMING [PART II]

While it is always difficult to optimize existing codes, an appropriate methodology can provide provable and replicable results. Last month, we saw that choosing the right test cases was essential to validate and evaluate optimizations, we then concentrate on efficiently using the compiler to optimize and to vectorize. This month, let's see how data layouts, algorithms implementation as well as parallelization impact performance...

ROMAIN DOLBEAU*

As previously mentioned, a minimal understanding of the hardware is required to exploit it. Efficiently utilizing the computational resources of the CPU is important, but since the term "memory wall" was coined by Wulf and McKee [1], it has been known that memory accesses were the limiting factor for many kinds of codes. The expression "bandwidth-bound algorithm" is dreaded by many of those tasked with improving performance, as it means that the available bandwidth from the main memory to the CPU cores is the limiting factor. But even without changing the algorithm, it is sometimes possible to drastically improve performance.

I - DATA LAYOUTS

One of the main issues with memory bandwidth is the amount wasted by poor data layouts. This issue is not new, and was already targeted by papers such as the work of Truong et al. [2]. But it still seems to be

* HPC Fellow, CAPS Entreprise

largely unknown to most developers. We first have to go back to computer architecture (and Hennessy and Patterson [3]). The main technique for a CPU to avoid the hundreds of cycles required for data to arrive from main memory are caches, tiny but fast memories close to the execution cores. While their existence is known, it seems that most programmers are unaware of their behavior and write code that greatly impedes their efficiency.

I.A - Arrays of structures vs. structures of arrays

Caches do not retain data at single element granularity, e.g. a single double-precision value. While this would help with temporal locality (the fact that a recently used element might soon be reused), it would do nothing for spatial locality (the fact that it's likely the next useful element will be a memory neighbor to a recently used element). To gain such spatial locality, caches have a granularity of a cache line, a contiguous amount of memory. Common sizes are 32 or 64 bytes of well-aligned memory (i.e. the address of the first byte is an integer multiple of the cache line size). Whenever a code requires an element from memory, the entire cache line is loaded from the main memory and retained in the cache. If a neighboring element from the same cache line is subsequently needed, the access will be a cache hit, and therefore very fast. This is a well known mechanism, taught in most computer science classes.

But the implications are not always well understood. If a cache line is 64 bytes, then every memory transaction will involve the whole 64 bytes, no matter how few bytes are actually needed by the code. If only a single double-precision element (8 bytes) is needed from each cache line, 87.5% of the memory bandwidth is wasted on unused bytes. Therefore, it is very important to ensure that as few elements as possible in each loaded cache line are not used. Unfortunately, some extremely common programming techniques goes against that principle. The most obvious offenders are structures (and of course objects, which are generally structures with associated functions).

While there is a lot to be said in favor of structures, that is not our subject; therefore we'll look at the downside, the way they can sometimes waste memory

References

[1] W. A. Wulf and S. A. McKee, *Hitting the memory wall: implications of the obvious*, SIGARCH Comput. Archit. News, vol. 23, no. 1, pp. 20–24, Mar. 1995.

[2] D. Truong, F. Bodin, and A. Seznec, *Improving cache behavior of dynamically allocated data structures*, in **Parallel Architectures and Compilation Techniques**, 1998. Proceedings. 1998 International Conference on, 1998, pp. 322–329.

[3] J. L. Hennessy and D. A. Patterson, **Computer Architecture: A Quantitative Approach**. San Mateo, CA: Morgan Kaufmann, 1990.

[4] A. V. Aho and J. E. Hopcroft, **The Design and Analysis of Computer Algorithms**, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1974.

[5] K. E. Atkinson, **An Introduction to Numerical Analysis**. New York: Wiley, 1978.

[6] S. Arora and B. Barak, **Computational Complexity: A Modern Approach**, 1st ed. New York, NY, USA: Cambridge University Press, 2009.

[7] E. Anderson, Z. Bai, J. Dongarra, A. Greenbaum, A. McKenney, J. Du Croz, S. Hammerling, J. Demmel, C. Bischof, and D. Sorensen, *Lapack: a portable linear algebra library for high-performance computers*, in **Proceedings of the 1990 ACM/IEEE conference on Supercomputing**, ser. Supercomputing '90. Los Alamitos, CA, USA: IEEE Computer Society Press, 1990, pp. 2–11.

.../...

bandwidth. The content of a well-defined structure is a lot of information related to an abstract concept in the code - it could be a particle used in fluid simulation, the current state of a point in a discretized space, or even an entire car. All the occurrences of that concept will be allocated in what is described as an Array of Structures. Subsequently, functions in the code will go through all occurrences in a loop to process the data, such as this:

```
foreach particle in
charged_particles
  update_velocity(particle,
  electric_field)
```

Presumably, the function **update_velocity** will change the speed of the charged particle in the electric field. But while the speed might be defined with only a few values (e.g. the current velocity in **X**, **Y** and **Z**), the structure itself is likely to contain much more information that's irrelevant to that particular step of computations. But as structures are contiguous in memory, much of that information will be loaded as they belong to the same cache line.

Let's assume each particle in our example is defined by 16 double-precision values, i.e. 128 bytes. The structure would occupy 2 full cache lines all by itself. Let's also assume all 3 velocities are in the same half of the structure, i.e. in the same cache line. Each time we need to update a velocity, the CPU will:

1 - Load the cache line (64 bytes) containing the three velocities (24 bytes);

2 - Perform any required computations, and update the value in the cache;

3 - Eventually, when space in the cache is needed, the whole cache line (64 bytes) will be flushed to memory.

62.5% of the memory bandwidth required to load and store the particle is wasted by unnecessary data. If the three

References (follow up)

[8] M. Frigo and S. G. Johnson, *The design and implementation of FFTW3*, **Proceedings of the IEEE**, vol. 93, no. 2, pp. 216–231, 2005, special issue on "Program Generation, Optimization, and Platform Adaptation".

[9] M. Frigo, *A fast Fourier transform compiler*, in **Proc. 1999 ACM SIGPLAN Conf. on Programming Language Design and Implementation**, vol. 34, no. 5. ACM, May 1999, pp. 169–180.

[10] M. Herlihy and N. Shavit, **The Art of Multiprocessor Programming**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.

[11] T. Mattson, B. Sanders, and B. Massingill, **Patterns for parallel programming**, 1st ed. Addison-Wesley Professional, 2004.

[12] B. Chapman, G. Jost, and R. v. d. Pas, **Using OpenMP: Portable Shared Memory Parallel Programming (Scientific and Engineering Computation)**. The MIT Press, 2007.

[13] W. Gropp, E. Lusk, and A. Skjellum, **Using MPI (2nd ed.): portable parallel programming with the message-passing interface**. Cambridge, MA, USA: MIT Press, 1999.

[14] A. S. Tanenbaum, **Modern operating systems**. Englewood Cliffs, N.J.: Prentice Hall, 1992.

[15] W. J. Bolosky and M. L. Scott, *False sharing and its effect on shared memory performance*, in **USENIX Systems** on USENIX Experiences with Distributed and Multiprocessor Systems - Volume 4, ser. Sedms'93. Berkeley, CA, USA: USENIX Association, 1993, pp. 3–3.

[16] D. E. Culler, A. Gupta, and J. P. Singh, **Parallel Computer Architecture: A Hardware/Software Approach**, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997.

velocities were spanning both halves of the structure, then the bandwidth requirement would double for the same amount of useful data - wasting 81.25%.

One solution for this issue is called structures of arrays. As the name implies, the idea is to inverse the relationship by first allocating all atoms of data in arrays, and then grouping them into a structure. Instead of having an array of particles, each with its velocities, the code would use three arrays of velocities. Each array would be contained as a single velocity in **X**, **Y** or **Z** for all particles. The result for the function **update_velocity** above is that we would need three cache lines instead of one: one for each of the **X**, **Y** and **Z** velocities. But that is not a bad thing. The first particle would require 192 bytes of bandwidth, loading 8 elements of each array. However, the next 7 particles would require no bandwidth at all - their data had been prefetched by the first particle, thanks to spatial locality. The average bandwidth per particle is therefore an optimal 24 bytes per particle, with a greatly reduced latency because of locality. If that function was purely bandwidth-bound, we would have gained a factor of x2.66 by reorganizing data in a cache-friendly manner.

I.B - Multidimensional array *vs* array of pointers

This is an issue that doesn't exist in Fortran, where multi-dimensional arrays are the norm. But in the C language family the issue is pervasive.

Listing 1 - Example of array of pointers.

```
void matrix_sum_aop(int n, double** A, double** B) {
  for (int i = 0 ; i < n ; i++) {
    for (int j = 0 ; j < n ; j++) {
      A[i][j] += B[i][j];
    }
  }
}
```

Listing 2 - Example of multidimensional arrays.

```
void matrix_sum_mda(int n, double A[n][n], double B[n][n]) {
  for (int i = 0 ; i < n ; i++) {
    for (int j = 0 ; j < n ; j++) {
      A[i][j] += B[i][j];
    }
  }
}
```

Listing 3 - Example of explicitly linearized arrays.

```
void matrix_sum_linear(int n, double* A, double* B) {
  for (int i = 0 ; i < n ; i++) {
    for (int j = 0 ; j < n ; j++) {
      A[i * n + j] += B[i * n + j];
    }
  }
}
```

A lot of code utilizes not multi-dimensional arrays but array of pointers, adding a useless intermediate load. **Listing 1** illustrates this form of inefficient code. The data is stored behind a pointer to a pointer (hence the two stars).

The body of the code looks good to the untrained eye: it has a nice pair of square brackets to access the element of the two dimensional data, as is done in Fortran. But the performance will indeed be suboptimal. The real meaning of the code includes not one, but two chained loads to access the data. The machine must first evaluate **A[i]**, itself a pointer to double. Then this pointer is used to retrieve **A[i][j]** i.e. the data itself. This is in effect an indirect access.

An efficient way can be exactly the same in the body, by utilizing a properly typed pointer to the data. This is illustrated in **listing 2**. By specifying the dimensions of the array in the parameter list, it becomes possible to use the clean multi-dimensional notation of the C language - which unfortunately looks exactly the same as a pointer-to-pointer double dereference. This new version will simply compute the linear access **i * n + j**, and do a single lookup in memory to access the

data. Of course, the data allocation is also different: instead of first allocating the array of the pointer, followed by a loop allocating all the pointers to double, a single allocation of the entire data set is used. The exact same allocation that would be used for the ugly, explicitly linearized version illustrated in **listing 3** that most programmers justifiably try to avoid.

II - ALGORITHMS

The word algorithm can cover different types of problems depending upon the audience. Computer scientists will think in terms of basic programming techniques as explained by Aho et al. [4]. Other scientists in need of high performance computing will likely think more in terms of numerical analysis as introduced by Atkinson [5].

In either cases, the idea is to achieve the desired results with minimal complexity, that is, with a minimal increase in work when augmenting the problem size. Complexity is usually taught in most computer science courses, and numerous recommendable books have been written on the subject such as Arora et al [6].

It is usually quite difficult for a computer scientist to replace a numerical algorithm chosen to solve a particular problem by a “better” one. This requires understanding what the code is trying to achieve in computer terms but also in physical (or chemical, astronomical...) terms as well. It also entails understanding the numerical stability, approximation, the boundary conditions, and whatever other requirements might exist for solving the underlying problem. This is something that must be done with involvement someone from the scientific field. More often than not, the constraints of numerical algorithms are such that they unfortunately cannot be changed.

What remains to be studied is the implementation of such algorithms, which quite often are built on top of other algorithms - those of the computer science variety. Matrix inversion, matrix multiplication, finding a maximum value, fast Fourier transforms and so on are the building blocks of many numerical algorithms. They are quite amenable to improvements, substitution and well-studied optimizations.

One of the primary tasks of the optimizer in a numerical code will be to identify these types of conventional algorithms, and make sure the implementation used is the best one for the code. This is usually not a lot of work, as vendor-supplied libraries will include most of them. For instance, the Intel Math Kernel Library (accessible in recent Intel compilers by simply calling the **-mkl** option in the compiler and the linker) includes full implementations of BLAS or LAPACK [7], fast Fourier transforms including a FFTW3 [8] compatible interface, and so on. One important aspect when using such libraries are their domain of efficiency: extremely small problem sizes are not well suited to the overhead of a specialized library. For instance, large matrix multiplications should make use of an optimized function such as **dgemm**. Small constant size matrix multiplications (such as 3×3 matrices used in Euclidean space rotation) should be kept as small hand-written functions, potentially amenable to aggressive inlining.

In a similar way, whereas large FFT can take advantage of FFTW3 itself or its MKL counterpart, small FFT of a given size can sometimes be readily implemented with the FFTW codelet generator described in [9]. A more work-intensive optimization is the specialization of algorithms for constant input data. For instance, if the small 3×3 matrix mentioned earlier is a rotation around the main axis of a 3D space, the presence of zeroes and one in the matrix can be hard-coded in the function, removing multiple useless operations.

III - PARALLEL CODES

Parallelization is one of the richest and most complex subjects of computer science, and is out of the scope of this paper. Many introductory books have been written on the subject among which for instance Herlihy & Shavit [10], Mattson, Sanders & Massingill [11], Chapman, Jost & Pas [12] and Gropp, Lusk & Skjellum [13]. This section deals with running parallel code on contemporary machines without falling into common pitfalls.

Figure 1 - Latency to walk through an array of varying size, form the lat_mem_rd benchmark.

III.A - NUMA

All recent multi-socket x86 64 systems are NUMA (Non-Uniform Memory Access), i.e. the latency of a load instruction depends on the address accessed by the load. The common implementation in all AMD Opterons and all Intel Xeons since the 55xx (i.e. Nehalem-EP family) is to have one or more memory controller(s) in each socket, and to connect each socket via a cache coherent dedicated network. In AMD systems the links are HyperTransport, while for Xeons they are QPI (QuickPath Interconnect). Whenever a CPU must access a memory address located in a memory chip connected to another socket, the request and result have to go through the network, adding additional latencies. Such access is therefore slower than accessing a memory address located in a locally connected memory chip. This effect is illustrated in **figure 1**, which plots the average latency of memory access when stepping through an array of varying size on a dual socket Xeon X5680 Westmere-EP. The first three horizontal plateaus are measurements where the array fits in one of the three levels of caches, while the much higher levels at over 12 MiB measure the time to access the main memory. The four lines represent the four possible relations between the computation and the memory: 0/0 and 1/1 represent access to locally attached memory, while 0/1 and 1/0 represent access to the other socket's memory. Going from about 65 ns to about 106 ns represents an approximately 63% increase in latency when accessing remote memory.

III.B - NUMA on Linux

To avoid these kinds of performance issues, it is important to understand how memory is allocated (virtually and physically) by the operating system, to ensure memory pages (the concept of paging is among those covered by Tanenbaum [14]) are physically located close to the CPU that will use them. The Linux page size will be 4 KiB or 2 MiB; the last one is more efficient for TLB (again [3]) but is not yet perfectly supported. This page size will be the granularity at which the memory can be placed in the available physical memories.

In Linux, this placement is made by specifying in which "NUMA node" the physical page should reside. The default behavior is only partially satisfactory. At the time the code requires allocation of memory, virtual space is reserved but no physical page is allocated. The first time an address inside the page is written, the page will be allocated on the same NUMA node that generated the write (if that memory is full, the system will fall back to allocating on other nodes). This is a very important aspect. If for one reason or another the memory is "touched" (written to) by only one thread, then all physical pages will live in the NUMA node where the thread was executing at the time.

Of course, if the advice from **first part** was followed, then the thread has been properly "pinned" on the CPU and will not move. This means than any subsequent access by this same thread will benefit from the minimal latency. That's the good aspect of the behavior.

One of the consequences is that in a two-socket system (two NUMA nodes), only half the memory will be used by the single-thread process (unless consuming a large amount of memory). It also means that only half the memory bandwidth from the entire system can be exploited, leaving the second memory controller in the second socket unused. Another consequence is that if the code subsequently becomes multi-threaded (for instance through the use of OpenMP directives), then half the threads will exclusively use remote memory, and performance will suffer accordingly.

Depending on the kind of workload, there are some simple heuristics to limit the nefarious effects of NUMA:

• **Single process, single thread**: the default behavior will only use the local NUMA node as explained. This might be a good thing (it optimizes the latency) or a bad thing (it cuts the available memory bandwidth in half). Some code might benefit from interleaving pages on all nodes, either by prefixing the command with **numactl --interleave=all**, or by replacing the allocation function such as malloc by a NUMA-specific function such as **numa_alloc_interleaved**.

• **Multiple processes, each single thread**: this is the norm for MPI-based code that does not include multithreading. The default behavior of Linux (provided the process have been properly "pinned" to their CPU) is excellent, as all processes will have low latency memory, yet the multiplicity of processes will ensure the use of all memory controllers and available bandwidth.

• **Single process, multiple threads**: this is the norm for OpenMP-based code. The default behavior is usually quite bad; more often than not, the code will cause the allocation of physical pages on the master thread's node, despite parallel sections running on all CPUs. Reading data from a file, receiving data from the network, explicit non-parallel zeroing or initialization of the allocated memory will all lead to this single-node allocation. Forcing interleaving as above usually helps (by utilizing all available bandwidth), but is not optimal (it forces the averaging of latency rather than minimizing it); it still should be tried first, as it is very easy to implement. The ideal solution is to be able to place each page on the node whose threads will make the most use of it, but that is not a simple task.

• **Multiple processes, with multiple threads**: although it might seem the most complicated case, it usually isn't: the typical placement of one process in each socket, with as many threads as there are cores in the socket leads to an optimal placement by the default behavior. No matter which thread allocates the memory, all the threads of the same process will see optimal latency.

III.C - False sharing

False sharing is a long-standing problem in a cache coherent multicore machine (see Bolosky and Scott [15]). Modern CPUs generally use invalidation-based coherency protocol (relevant details about it can be found in Culler et al. [16]), whereas every write to a memory address requires the corresponding cache line to be present in the local CPU cache with exclusive write access-that is, no other cache holds a valid copy.

"True sharing" occurs when a single data element is used simultaneously by more than one CPU. It requires careful synchronization to ensure respect of the semantic, and incurs a performance penalty as each update requires the invalidation of the other CPU's copy. Back-and-forth can become very costly but is unavoidable if required by the algorithm (unless, of course, one changes the algorithm).

"False sharing" occurs when two different data elements are used by two different CPUs, but happen to reside in the same cache line. There is no need for synchronization. However, because the granularity of invalidation is the whole cache line, each data update by a CPU will invalidate the other CPU's copy. This requires costly back-and-forth of the cache line. This can greatly affect performance and is completely unnecessary: if the two data elements lived in different cache lines, no invalidation would occur for either, thus ensuring maximum performance.

For very small data structures (such as a per-thread counter), this can be a huge problem as each update will lead to an unnecessary invalidation. For instance, **listing 4** describes a structure of two elements: **neg** and **pos**. Without the intermediate padding array pad, those two values would share a cache line, and when used by two different threads, would cause a major performance issue. The padding ensures that they are in different cache lines, thus avoiding the problem. A synthetic benchmark using this structure running two threads on two different sockets in a dual X5680 system would see a increase in time of 50% as opposed to when the padding was not in use. Hardware counters show a very large number of cache line replacement in the modified state when padding is not used, versus a negligible number when it is (cache coherency and the modified state are well explained by Culler et al. [16]).

For large arrays that are updated in parallel by multiple threads, the data distribution among threads will be an important factor. If using a round-robin distribution (i.e. thread number n out of N total threads updates all elements of indices **m** such as **n --- m (mod N))**, then false sharing will occur on most updates - not good. But if a block distribution was used, where each thread accesses 1/N of the elements in a single continuous block, then false sharing only occurs on cache lines at the boundaries of each block - at most one cache line will have shared data between two threads. Not only will the number of unnecessary invalidations be small relative to the total number of accesses, but they will usually be masked by the fact that one thread will update at the beginning of the computation, and the other one at the end, thus minimizing the effect. And to ensure conflict-free accesses, the block size that each thread handles can be rounded to a integer multiple of the cache line size,

Listing 4 - False sharing.

```
typedef struct {
  int neg;
#ifdef NOCONFLICT
  int pad[15];
#endif
  int pos;
} counters;
```

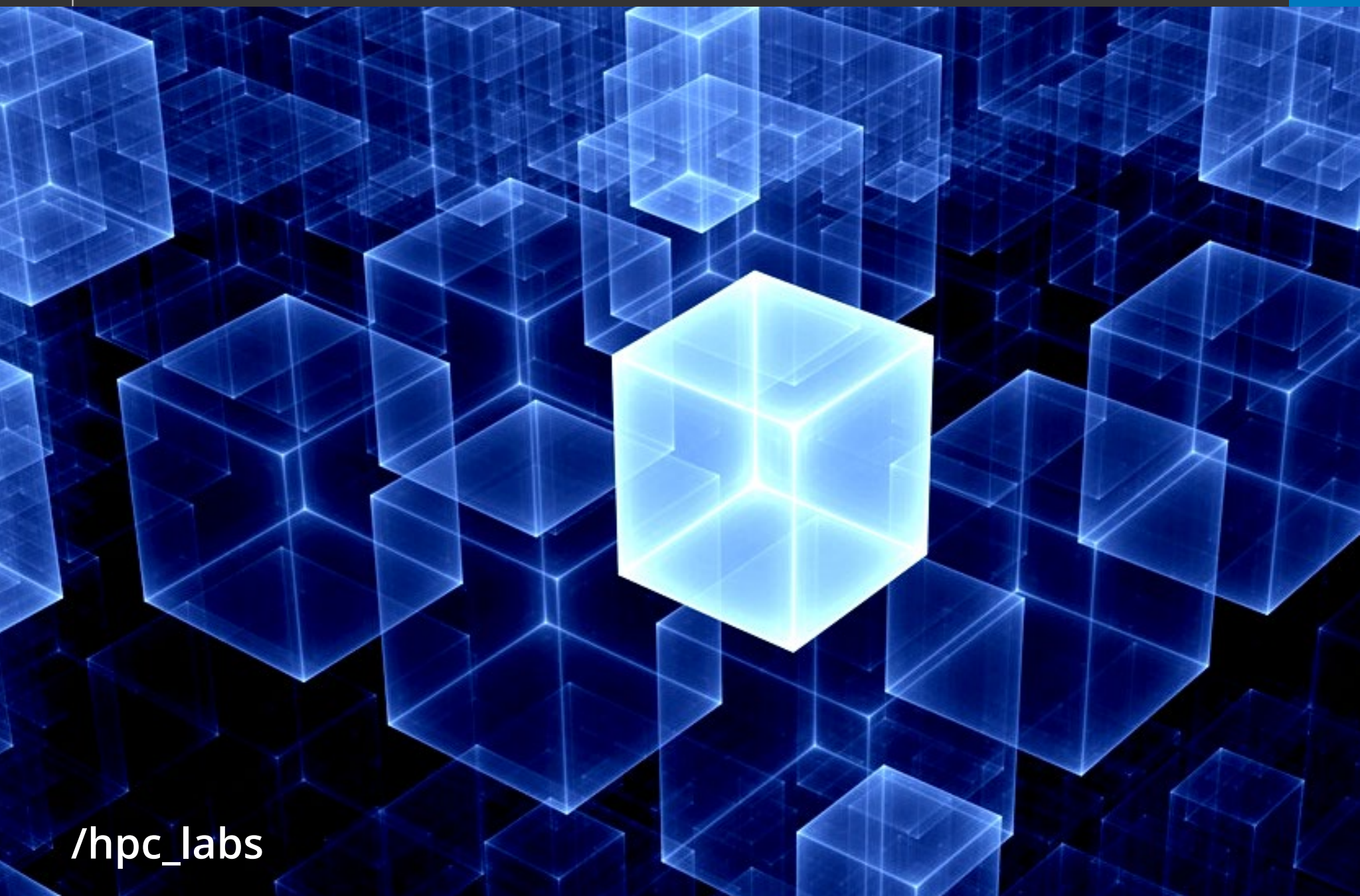
at the cost of a slightly less efficient load balancing between threads.

CONCLUSION

Here is the end of this series of two articles that introduces a few key points for the newcomer to keep in mind when trying to improve the performance of code. To summarize, it is important to be able to justify the validity of the work done, in terms of reliability, speed and the choice made when improving the code. Other points are to properly exploit the tools at hand before modifying the code, to understand the relation between the data structures and the performance of the code, avoid re-inventing the wheel and exploit pre-existing high-performance implementations of algorithms, and finally run parallel code in a manner adapted to the underlying hardware.

Hopefully, this will help programmers and scientists alike to understand some key aspects of performance and obtain higher performing code without an excessive amount of work.

Happy programming! ■



/hpc_labs

DISCOVERING OpenSPL, THE SPATIAL PROGRAMMING LANGUAGE

Behind OpenSPL is the factual concern that the temporal computation model of multicore processors reaches its limits in terms of scalability, efficiency, and complexity. This requires radical innovations in the design of systems destined for scale-up. What does spatial mean here? How OpenSPL works and what is the future of this new open programming language? Let's have a deep view on these...

OSKAR MENCER[1], MICHAEL J. FLYNN[2], JOHN P. SHEN[3]

OpenSPL is a novel initiative by CME Group, Chevron, Juniper and Maxeler Technologies to increase awareness and acceptability of computing in space rather than in time sequence.

While the initiative is new, the concepts and ideas behind computing in space have been used in practice for a very long time. In essence, making an Application Specific Integrated Circuit (ASIC) is just like computing in space where writing the program takes many hundreds of engineers and several years of effort, even without considering the enormous costs. As such, a single ASIC is now so expensive that it has to be able to execute all possible programs and applications. Computing in space, on the other hand, brings the silicon substrate to

(1) CEO Maxeler Technologies.
(2) Professor, Stanford University.
(3) Head of Nokia Research Center North America Lab.

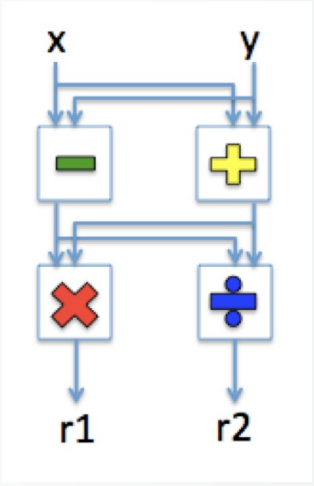
the trial-and-error programmer and allows us to adapt the spatial structure to the problem the substrate is solving.

Computing in space could be considered a generalization of decades of research (including the authors' contributions) towards systolic arrays, vector supercomputers, and a wide range of research projects such as Alan Huang's MIT Thesis on Computational Origami [1]. Going further back, to the late 1950s, IBM revolutionized hardware design with its Standard Modular System (SMS) using standardized circuit cards that could be manufactured quicker and more reliably than the older custom approach. A key piece of SMS was Automated Logic Design (ALD) sheets. The designer used a sheet with 2D grid of blocks; each block specified a logic function (card) and a connection. Each connected block was entered on a punched card, enabling the design to be managed by computer. The "compiled" design checked the logic, updated signal data and produced the wire routing for manufacturing.

This automated design and manufacturing process lead to the market dominance of IBM 1401 and 7090, among other machines The SMS was a precursor to the SLT design system used in System 360 and the age of the mainframe. OpenSPL-based systems gain a similar advantage from employing the lessons of manufacturing (and early IBM systems) to using assembly lines to build the results of computation.

Listing 1 - An OpenSPL kernel using Java syntax.

```
SCSVar x = io.input("x");
SCSVar y = io.input("y");
SCSVar t1 = x - y;
SCSVar t2 = x + y;
SCSVar r1 = t1 * t2;
SCSVar r2 = t1 / t2;
io.output ("r1", r1);
io.output ("r2", r2);
```



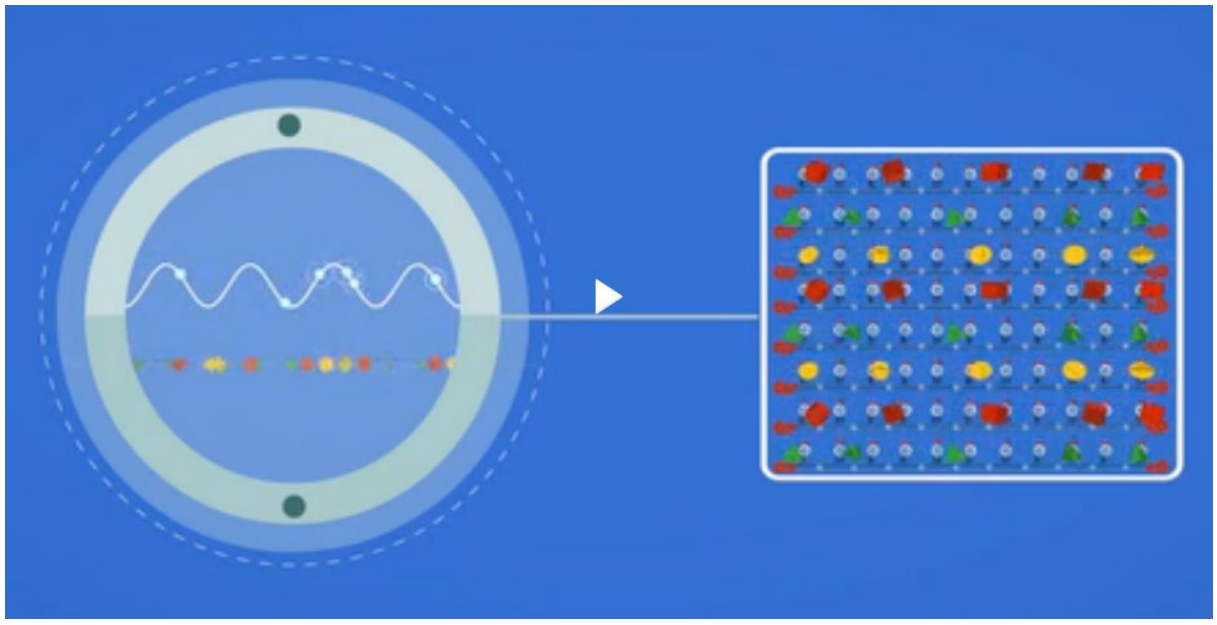


Fig. 1: OpenSPL split of computation into control flow and dataflow.

Back to 2014, what does OpenSPL mean in practice? Rather than describing a thread of instructions and duplicating the thread onto multiple processors to handle multiple streams of data (SIMD), OpenSPL requires a split of the computation into control flow and dataflow components, creating a network of arithmetic units (see **Figure 1**) through which the data flows just like the materials flow through a network of assembly lines in a modern factory. Each arithmetic unit forwards the result of the operation to the next arithmetic unit, eliminating the need for most register and memory accesses.

How does a spatial program work? Easy: describe a directed graph with sources and sinks, and connect sources and sinks to memory and other I/O channels. This can be done by using the syntax of any programming language. Maybe we could have called OpenSPL computing with directed graphs, but that seemed less elegant. The **listing 1** illustrates this.

Here SCS stands for Spatial Computing Substrate and hints at the purpose of the OpenSPL initiative. Rather than pretend that a single program could optimally compile to different architectures and substrates, we set OpenSPL as a baseline

[1] *The folding of circuits and systems,* Applied Optics, 31-26, 1992.

from which members of the initiative can construct substrate-specific compilers and runtime implementations.

The most frequent question at this point is: How about conditionals? The answer is that "conditionals" or "if" statements are not all created equal. There are really three types of conditionals that any programmer can distinguish, but that are hard for machines or compilers to reverse engineer as they look at the code: (1) conditional assignment, (2) data forks, and (3) separate global paths through a program.

Computing in Space has three separate natural mechanisms to deal with the three types of conditionals: (1) is a simple multiplexor allowing the programmer to select between two data producers. (2) is a fork in one graph or a fork driving data to two different graphs at the same time. (3) requires a bit of work in separating the different global paths into separate code and implementation. The code reorganization required by (3) is the most unpleasant part for the modern programmer, especially ones trained in the art of C++. Luckily, the untangling of global paths only has to be done to the parts of the program that take most of the time, which are typically small parts.

So now that we can program in 2D space, we need actual computers to do this on. The standard describes the concept of spatial computing substrates, or architectures that

Listing 2 - Simple Controlflow in Dataflow Kernel example.

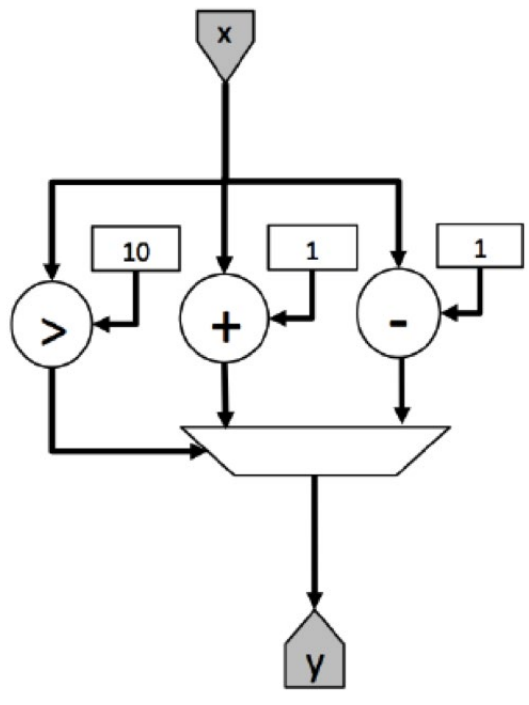
Both candidates (x+1 and x-1) are simultaneously computed in space and only the correct value flows out via the SCSVar result.

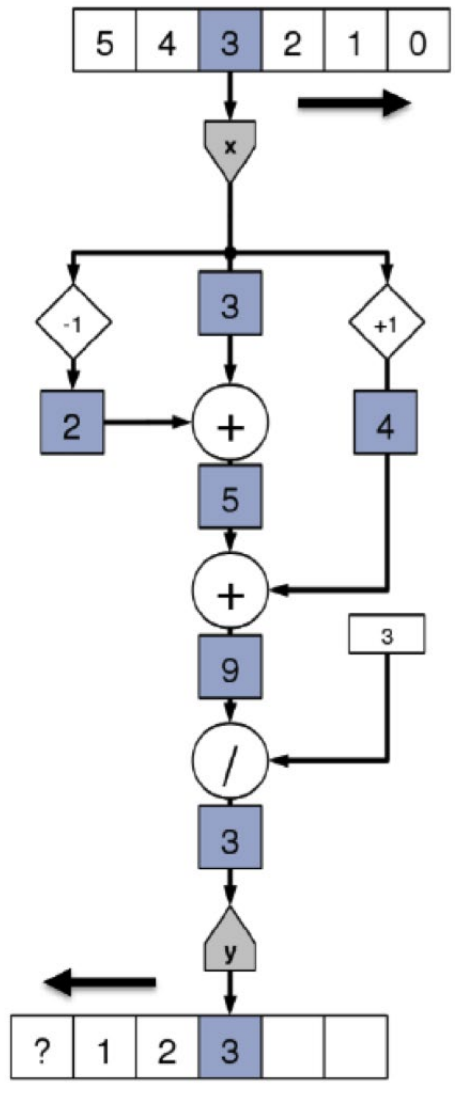
```
class Simple Kernel extends Kernel {
  SimpleKernel() {
    SCSVar x = io.input("x", scsFix(24));
    SCSVar result = (x > 10) ? x+1 : x--1;
    io.output("y", result, scsFix(25));
  }
}
```

Listing 3 - Moving Average Kernel example.

Application specific floating point precision numbers with 7-bit exponent and 17-bit mantissa are used.

```
class MovingAvgSimpleKernel extends Kernel {
  MovingAvgSimpleKernel() {
    SCSVar x = io.input("x", scsFloat(7, 17));
    SCSVar prev = stream.offset(x, --1);
    SCSVar next = stream.offset(x, 1);
    SCSVar sum = prev + x + next;
    SCSVar result = (sum / 3);
    io.output("y", result, scsFloat(7, 17));
  }
}
```





The Kernel graphs of the two examples. All coefficients (10, 1 and 1 on the left and 3 on the right) can be set externally using the SCS IO interface. In the right graph integer, data values are used instead of the example's custom floating point (7, 17) for the sake of simplicity.

support the spatial computing paradigm. While the standard is in its infancy, and is still being refined, we already have a first commercially available substrate: Multiscale Dataflow Computing by Maxeler Technologies. With a range of server, networking and soon also storage products for high end enterprise-level computing, such a substrate can be applied to a wide range of application domains and computing activities.

Multiscale Dataflow Computing already comes with simulators, debuggers and a university program having brought access to the technology to over 100 universities worldwide. As the standard continues developing, there could be many new Substrates for Computing in Space. Such new substrates could also bring an order of magnitude reduction in power consumption and physical size of the device to a wide range of additional domains such as wearable computing, embedded devices, mobile terminals and smart dust computing.

One promising domain is Mobile Supercomputing. Leveraging the OpenSPL standard and a new substrate for real-time sensing and processing, mobile supercomputing systems with order of magnitude reduction in both power consumption and physical system size can become feasible. Such mobile supercomputing systems will become essential in the emerging Mobile Computing Universe.

The emerging Mobile Computing Universe consists of the cloud infrastructure, personal mobile devices, and embedded environmental sensors (or "Internet of Things"). To support the real time processing of massive amounts of mobile data and the performing of deep analysis and inference to extract value from such big data, Exascale supercomputing infrastructure will be needed. However, due to the massive and global scale of this universe, implementing centralized cloud infrastructure to support such real-time supercomputing is not the most efficient approach. Distributing Petascale mobile supercomputers to the edge of the cloud is much more efficient and provides much better service.

The OpenSPL standard based on the dataflow computing model with appropriate substrate supported by powerful software tools offers the potential of achieving two orders of magnitude improvement on the performance/power ratio, and the opportunity to build mobile supercomputer systems that can be widely deployed without requiring special machine rooms and the associated supporting infrastructure. There is a real opportunity to bring power and energy efficient supercomputing to the mobile mass market. ■

The MPC-X1000 providing eight dataflow engines.

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