Report from DeIC SC14 Fact Finding tour

In November 2014 representatives from 4 Danish Universities (KU, DTU, SDU, AU) participated in the DeiC-sponsored Fact Finding tour to the SC14-conference in New Orleans. Before the conference, most of us visited Dell's Research Labs in Round Rock, Texas, and TACC, the Texas Advanced Computing Center, University of Texas at Austin. We also attended the Intel HPC Developer Conference in New Orleans. During the conference several private meetings were held with vendors of HPC equipment.

The aim of this report is briefly to report the group's findings. Please observe that a major part of our meetings were held under Non Disclosure Agreements (NDA), implying that we cannot reveal issues discussed or information shown to us during these meetings.

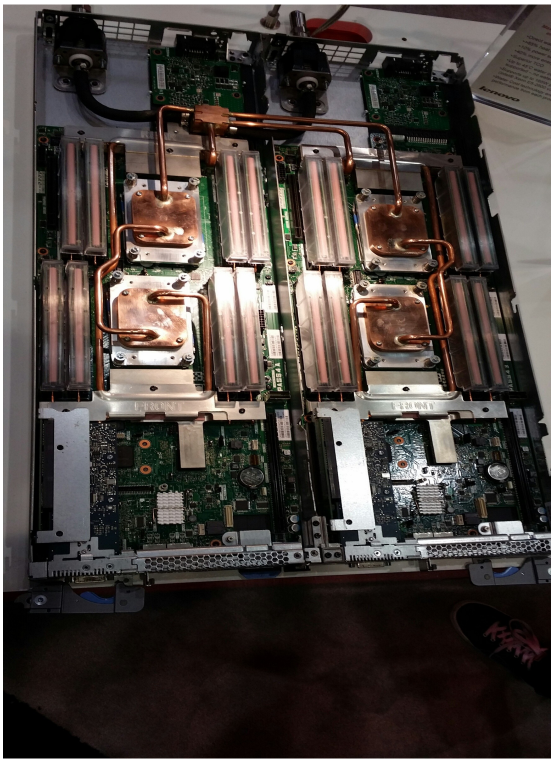
Technology update:

1. CPU processor update:
   - Intel Haswell performance depends significantly on the usage of AVX instructions, temperature, and chip samples within the same SKU. Higher memory performance is obtained with 2 memory controllers in the 10+ CPU core SKUs.
   - Power8, Power8+, Power9 / OpenPower. Linux OSes (not AIX) will be available.
   - AMD server processor road-map was presented at a private meeting.
   - ARM: 4 core 64-bit CPU
2. Accelerator update:
   - Xeon Phi “Knights Landing” will arrive in 2H 2015 in both bus and socket versions. The socket version will boot standard OSes and be binary compatible with Xeon. Network interconnect will be integrated.
   - Nvidia updates on Kepler, Pascal and Volta GPUs. Nvlink will be used as interconnect. Support for standard programming languages such as Java, Python, OpenACC, PGI compilers, Matlab.
   - AMD updates on “Firepro” and “Greenland”. PathScale compiler support for OpenMP and Python.
3. Interconnect update:
   - Mellanox Infiniband is moving to EDR speed.
   - Ethernet speed available at 10, 40 and 100 Gbps.
4. Server form factors continue to evolve significantly from all major vendors towards ever more compact cabinets. Solutions with 2 servers in 1U have become standard, and even more compact servers with 3 or 4 servers in 1U have been announced. A 4-GPU accelerator model in a 1U server has been announced by Dell.
5. Storage is, as in the rest of the industry, moving towards *Software Defined Storage* (SDS). In the HPC world we have long known that storage is mostly about software when moving up the stack to SANs and shared filesystems. What is new is that hardware vendors are embracing both open and closed SDS solutions. This makes it possible to avoid vendor lock-in on the hardware side of storage. Interesting new trends are:
   - IBM have long been in the HPC storage business with GPFS. They have jumped on the SDS bandwagon with the *IBM Elastic Storage* re-brand. However, you need to get IBM hardware to get all of the features. This should also be available from Lenovo.
   - IBM has also moved into new markets such as Big Data, and the same is true for Isilon. While Isilon doesn't offer a traditional HPC filesystem (not parallel, no Infiniband connection from clients), they are moving into Big Data.
   - Hardware vendors are starting to offer and support certified solutions on top of open solutions such as Lustre, GlusterFS and Ceph. With Lustre

this is mainly through Intel® Enterprise Edition for Lustre, while GlusterFS and Ceph products are up and coming throughout the storage industry. RedHat is the main driving factor behind this, as they purchased the companies that developed (open-source) GlusterFS and Ceph.

- At the Seagate presentation we got a great look into the future of spinning disks. Obviously there will be higher data density, but Seagate is also working on making disk I/O faster. This is very important when 20TB disks will arrive in the not so distant future.
- We had a meeting with the vendor of BeeGFS. Storage replication is probably the most interesting feature coming from them. AU has a system in production with BeeGFS – in fact, it is the largest BeeGFS system in the world! - and part of the meeting was about implementing a new feature in BeeGFS to support their needs. It was very interesting to see how open-minded the people from BeeGFS are to such requests.
- Many hardware vendors showed new storage chassis with high disk densities. The most impressive was a 4U chassis from Dell that could fit 90 3.5” drives and two full dual-socket servers in the back. With 10+ TB disks it will become possible to reach 10PB of capacity in a single rack.

6. Middleware/Scheduler. The major resource manager products are currently SLURM, Moab and LSF. SLURM is used on 6 of the top 10 TOP500 supercomputers, and is available as Open Source with optional support. LSF and Moab are commercially sold products, LSF is owned by IBM.
7. Liquid cooling is available from major vendors, but the technologies have not converged towards similar solutions. Cooling water of up to 45 degrees may be used either in the rack mid-plane, directly on processors and DIMMs, or in water cooled back doors. Closed loop heat exchangers must be used, and water quality must be monitored constantly for the risks of leakage or bio-growth. Recycling of waste heat is used in a few data centers.
   Liquid immersion cooling, which is available from several vendors, is being evaluated at some HPC centers, for example, TACC.
8. Datacenter
   - Power density per rack continues to increase with ever more compact servers and CPU and GPU power increases. Depending on density and GPU configuration, a single rack may use 30, 40, 50 or even 60 kW of power. Innovative cooling solutions must be implemented in high-power racks.
   - Container solutions. No containers were on display at SC14, but we had a tour of the Dell container in Round Rock. Containers are complex solutions requiring many innovations.



Two IBM NextScale dual-CPU servers in one 1U chassis on display in the SC14 exhibition hall.
All CPUs and memory DIMMs are water cooled.

Announcements:

1. Top500 – for the first time we see a declining trend in the growth of performance. By carefully analyzing the numbers from the list, it turns out that the CPU technology isn't causing the flattening of the curve. Apparently the accelerators are part of the problem.

   Selected entries from the November 2014 TOP500 list:

Rank on TOP500	Name	Country	Performance
1	Tianhe-2	China	33.8 PFlops
2	Titan	ORNL, USA	17.6 PFlops
3	Sequoia	LLNL, USA	17.1 PFlops
6	Piz Daint	CSCS, CH	6.27 PFlops
7	Stampede	TACC, USA	5.17 PFlops
121	Computerome	DTU-CBS, DK	0.411 PFlops
446	Vestas	Vestas, DK	0.162 PFlops
500	“cluster”	USA	0.153 PFlops

   No. 6 on the list is the fastest computer in Europe.
   No. 121 and 446 on the list are the fastest (officially known) computers in Denmark.

2. Two new Power9-based supercomputers at ORNL (“Summit”) and LLNL (“Sierra”) will be put into operation during 2017. Both systems will be equipped with Nvidia accelerators (Volta series) and Mellanox EDR Infiniband. These cards will be utilizing a new faster bus, Nvlink, instead of PCIe.
3. Mellanox is now launching 100 Gbps Ethernet and Infiniband (EDR).
4. Other vendors are also shipping 100 Gbps Ethernet switches (for example Brocade, Arista, Cisco).
5. Nvidia is now shipping the K80 GPU – a twin K40 GPU.
6. New interesting CUDA programming paradigms for the next generation of Nvidia GPUs, allowing for an easier way to deal with data copying between GPU- and CPU- memory via the notion of virtual GPU-memory.
7. AMD announced that they want to re-enter the HPC market with new interesting GPU and CPU models. The time frame is 2-3 years from now.
8. IBM has sold their x86-business to Lenovo. In Denmark this will become reality from January 1st 2015.

Visits:

- Dell Research Labs, Round Rock.
- TACC, Texas Advanced Computing Center, University of Texas at Austin.
  - Home of the *Stampede* supercomputer, which is no. 7 on the current Top500 list with a performance of 5.1 PFlops.
  - TACC operates its own power plant providing the necessary 4.5 MW.
  - 90% of the computing power is for users holding an NSF grant,
5% is reserved for users at University of Texas, and
5% is for TACC usage, for example, to be sold to commercial customers.
  - Stampede uses the *Slurm* scheduler.
  - TACC employs about 20 consultants, most of them fully paid by TACC, to maintain discipline-specific program packages and provide user support. Most of the consultants are also active researchers.
  - No special security setups are made to satisfy any possible needs from commercial customers.
  - *Stampede* has one scheduled maintenance day per month.
  - TACC uses Lustre for all shared filesystems.
- Intel HPC Developer Conference:
  - Presentation of CPU and accelerator roadmaps.
- SC14 Conference:
  - Technical Program (talks, BOFs, poster sessions).
  - Exhibition show floor.

Other findings of interest:

- Nvidia: Offers to help and support customers (f.ex. via DeiC's Kompetancecenter?)