DEIC REPORT 2017

Fact Finding Tour at Super Computing 17, Nov. Denver, Colorado

Boye, Mads - mb@its.aau.dk
Hansen, Niels Carl - ncwh@phys.au.dk
Happe, Hans Henrik - happe@nbi.ku.dk
Nielsen, Ole Holm - ole.h.nielsen@fysik.dtu.dk
Madsen, Erik B. - erikm@sdu.dk
Madsen, Torben K. - tm@sdu.dk
Visling, Jannick - janvi@sdu.dk

2017-12-15

Contents

Preface

The SC17 International Conference for High Performance Computing, Networking, Storage and Analysis was held in Denver, Colorado where a delegation represented the five Danish universities:

Aarhus University (AU), Technical University of Denmark (DTU), University of Copenhagen (KU), University of Southern Denmark (SDU), and Aalborg University (AAU).

The purpose of the present document is to briefly report the delegation's findings. The trip was subsidized in part by the Danish e-infrastructure Cooperation (DeiC). Prior to SC17 the delegation participated in the HP-CAST 29 conference and the INTEL HPC Developer Conference 2017. During these conferences the delegation attended a number of Birds of a Feather sessions (BoF), Talks, and Sessions. In addition, the delegation had arranged meetings with a number of vendors to gain knowledge of hardware and software developments within the HPC field. Information obtained during vendor meetings was mostly under Non-disclosure agreements (NDA), and such information can not be disclosed in this report.

Central Processing Unit (CPU)

AMD

AMD seems to be back in the market. Armed with new EPYC 7000 Zen series they aim for the data center. Their first Zen CPU codename Naples was released in June 2017. AMD currently provide 9 models, with the following specs[5]:

- Number of cores: 8,16,24,32 (14nm).
- Base Clock Speed 2.0Ghz 2.4GHz.
- Max Turbo Core Speed: 2.7GHz 3.2GHz.
- Default TDP: 120W 180W.
- Max PCI-E 3.0 lanes: 128.
- 8-channel memory architecture.
- Max SATA3: 32.
- Max NVMe: 32.

The CPU can be 1-socket models and 2-socket models. The 1-socket models can be identified by the *P* postfix. These only support 1-socket, and can not be used in a 2-socket board. The 1p model cost less. In a none scalable system they would be a good choice. The high number of PCIe lanes makes the ADM and interesting choice. Also at low cost and power consumption, it seems like AMD can take on Intel. Various hardware vendors, such as Dell, HPE and Lenovo, are expected to supply AMD based systems.

Futre road map fo the EPYC CPU was told under NDA, and can therefore not be disclosed.

ARM

The chip maker company Cavium presented the ThunderX2 ARMv8 CPU at their booth. ThunderX2 is a impressing CPU targeted for data center applications (virtualizations, web services, etc). Some of the key specs are:

- up to 32 physical cores per CPU.
- support for one- or 2-socket systems.
- 8 DDR4 memory channels - max. 16 DDR4 DIMMs pr. socket.
- PCIe gen3, up to 108 lanes in a 2-socket configuration.
- Power consumption assumed to be about 195W/CPU.

More manufactures plans to equip servers with ThunderX2 CPUs, e.g. HPE in the Apollo 70 system.

Intel

Intel currently ships the mainstream Xeon "Skylake" processors, among which the "Platinum" and "Gold" series are most relevant for HPC. Intel's plans for the future Xeon processor generations Cascade Lake, Ice Lake, and Xeon Phi accelerators, were described under NDA.

Accelerators

AMD

AMD plan to relaunch their GPU series at some point during 2018. These will be 100% open source and focused on machine intelligence and deep learning. First card available will be the Radeon Instinct MI25[6].

Intel

Intel brings little news with plans to their accelerators, but will probably be revised based on recent customer and overall market needs. There are rumors[11] on the Internet about the future of the Xeon Phi product line, but nothing official

Intel will focus development on Knights Mill which is built on a completely different architecture specialized to accelerate deep learning. Details on Knights Mill are still limited. A guess is that it will be based on technology from the AI processor vendor Nervana, which was acquired by Intel in late 2016. Intel is also working on FPGA based accelerators, based on technology from Altera, which Intel acquired a few years ago. The FPGA is not ready for mainstream HPC usage.

Nvidia

Nvidia has had a successful launch of the new Tesla V100 card. There was no new hardware announcements at this years SC17. Nvidia has committed to update many of the bigger deep learning software stacks to work with the latest Nvidia GPU, to ensure the best possible performance.

Nvidia announced that they in collaboration with Amazon provides the first Tesla V100 compute cloud, where cloud servers can be allocated with up to eight Tesla V100 GPU's.

Flash Technologies

3D-xPoint/NVMe/Apache Pass

The 3D-XPoint technology is now available, but only in the form of Optane NVMe drives. The Apache Pass DIMM modules is still not available, but the launch is getting closer. It should however be possible to use the Optane NVME to complement the systems DRAM, at least when it comes to non-memory bound workloads.

Via an Intel supplied middleware layer, expand the system memory of a 2-socket system up to 24TB, with cheap, compared to DRAM, Optane NVMe. Performance of the NVMe's are around 1/20th of DRAM but the middleware will try to compensate by prefetching to real DRAM. This middleware can also boost performance of non-NUMA aware applications, as it is capable of moving data around, so that it is closer to the processor that is using it.

Interconnect/Fabric

Gen-Z

The evolution of fabrics and topologies Using HPC interconnects of today, which operate in the range of 50 to 100 pJ per bit, energizing an exascale systems fabric alone would consume more than the Department of Energy's (DoE) 20-megawatt power target. Therefore a physical signaling layer that is an order of magnitude more efficient, must be designed and built.

Hewlett Packard Enterprise, other companies, and labs are aiming to deliver the highly desired target of 1 pJ/bit within the next decade, through research and development in a very energetic optical technology, based off vertical cavity surface-emitting lasers (VCSELs) and silicon photonics.

HPE is working at an eight VCSEL electrical-to-optical transceiver modules on each fabric switch board with each optical module operating at a 1.2 Tb/sec bidirectional data rate. The aggregate fabric switch bandwidth of each board is 9.6 Tb/sec with an aggregate speed of 19.2 Tb/sec of bandwidth per chassis.

Mellanox

Ahead of SC17, Mellanox announced a new line of scalable switches, to provide higher flexibility in terms of numbers of switch in the fabric,e.g. up to 1.600 port of 100Gb/s. Mellanox had some interesting improvements to handle mixed version of InfiniBand in the same subnet.

On the software side, Mellanox also have some interesting improvements. Introducing "Scalable Hierachical Aggregation and Reduction Protocol" (SHARP), which is capable of performance optimizing the MPI workloads, but offloading computations to the network, and thereby reducing computation time. This should be handled by auto discovery. another introduction is the "Self-Healing Interconnect Technology" (SHEILD), which makes the InfiniBand network more resilience towards failures, as the network automatically can detect errors, and re-route the job, instead of letting it crash.

Omni-Path

The current Intel Omni-Path fabric competes with Mellanox products at 100 Gbit/s speed. Host adapters and switches differ in functionality between these two main vendors. Future plans for the Omni-Path fabric, its backwards compatibility, as well as the Xeon -F processors with integrated Omni-Path were described under NDA.

Servers

Dell EMC

Dell EMC will release servers with the AMD EPYC processor soon. There are hints on the Internet already. A 1-socket 1U and a 2-socket 2U server. Also rumors about a 1-socket 2U server.

Dell EMC released their Skylake-SP upgrade of the C4130 GPU server during the conference. It is called C4140 and can still have 4x GPUs in a 1U form factor. Also, the new Nvidia V100 GPUs are supported. Both the socket version (SXM2) and the PCI-e version.

Dell showed a liquid cooled C6420, which is based on CoolIT technology[1]. Dell are now able to deliver server fitted with liquid cooling out of the factory. The delegation agreed that it was easy to bend the hoses out of shape, which might result in a leak.

Huawei

The delegation had an NDA meeting with the Director of Product Management from server vendor Huawei. The Huawei HPC server portfolio was described, including support processors from Intel and ARM, as well as GPU and FPGA accelerators. On the SC17 show floor the E9000 16-blade chassis, the X6000 4-node chassis, and the FusionServer G560, an eight GPU servers solution, were demonstrated.

HPE

HPE is upgrading most of their HPC portfolio to Gen 10, which means that the Apollo line now supports Intel Xeon Scaleable Processors. In addition the Apollo 2000 chassis does now support Nvidia Tesla V100.

The amount of disks in Apollo 4510 gen 10 has been scaled down to 60 disks, as they change the design to disk drawers instead, which makes the disks easier service. Under NDA the was informed of HPE's plans towards ARM and AMD on both the ProLiant and Apollo server platform.

Lenovo

Known for the wide range of product from desktops to HPC, Lenovo has a lot of new products planned, but due to the NDA, the roadmap and new products, and adoptation of new hardware, cannot be disclosed in this report.

Lenovo continues to develop their water-cooling, and are using copper tubing which they claim have tested for 7 years without leakage. Test with aluminum tubing resulted in leakage after 2-3 years[7].

Storage

BeeGFS

The main question about BeeGFS is what the difference is to Lustre? Some of the improvements you will have with BeeGFS is:

- Metadata servers uses buddies as a Raid/Mirror if one goes down, the other will be available.

BeeGFS is all open source. However, the team will review all suggestions/changes before adding it to the main code. The latest version is 7. Though this is still just a release candidate it should be available within a few weeks. Some of the key elements in the new release are:

- Storage pools.
- Modifications in the event log.
- BeeGFS-Man (preview, stable but incomplete).

Installation is fairly easy, which more users in the audience at their BoF could acknowledge. The performance tuning is more complicated, compared with turnkey solutions, various guidelines is provided at the BeeGFS webpage. A Windows client is in the making for release about Q2 2018[8].

CEPH

Prior to SC17 a new version of CEPH was released, codename Luminous, which is an LTS. The CEPH BoF described all the new features that is now available in CEPH. The release schedule changes from every 6th month, to every 0th month. The next release will be mimic. Highlights of the new features is listed below:

- Web dashboard fro monitoring cluster status.
- BlueStore - new backend for ceph-ods. BlueStore provides new direct handling of HDDS in ods nodes, with out an intermediate file system, like XFS.
- Support for full data and metedata checksums of all data stored.

- Inline file compression.
- Added support for erasure coding.
- New daemon named *ceph-mgr*. The new daemon will handle metric related calls. A new REST API has been added.
- Added better auto tuning of OSD's, based on underlying hardware (HDD or SSD)
- Added support for InfiniBand for frontend data access.
- Support for multiple meta data servers.

Many larger universities now has large CEPH installation, in the 2PB+ range. Depending on configuration and workload, mostly read workloads, some users does see better performance than on file systems like Lustre og BeeGFS.

Lustre

This spring it was announced that Intel discontinues their Intel branded releases of Lustre. However, they will contribute directly to the open-source project. They will still provide level 3 support[2]. Furthermore, Intel open sourced their management software, Intel Management for Lustre (IML). This does not change much for the big turnkey Lustre suppliers like DDN, Cray and SGI (HPE). They already provided level 1 and 2 support on their own. They were all represented with Lustre solutions at SC'17. For self-supporting Lustre users this change is welcomed, as users will get long-term releases and maintenance releases between major releases.

The Lustre Community BoF was well attended. There was a panel of developers, mainly from Intel and a presentation about the road map was given[3]. To name a few highlights from the road map:

- Data-on-MDT: Small file data can be stored on the meta data servers as a top tier. This gives small file creation lower latency and more IOPS.
- File level redundancy: Makes it possible to define redundancy on a per file basis. Policy is defined per directory, so new files an directories below will inherit. Mirroring to start with, while erasure coding are planned.

Data centers / Server Rooms

Water/Liquid Cooling

The power density and therefore the heat exchange in HPC products is steadily increasing. In addition to standard air-cooled solutions, all HPC server vendors now offer liquid cooling solutions, but with quite different approaches to direct liquid cooling of processors inside the server. Some vendors can liquid-cool also DIMM-memory, GPUs, and the network fabric adapters.

Another approach is liquid cooling of fully contained racks, where the servers are air-cooled. Some solutions offer 90%+ heat removal, giving a minimal heat impact on the surroundings.

The liquid distribution inside servers comes in many forms. Some vendors have decided to focus on copper tubing for reliability e.g. Lenovo, while others use plastic or rubber piping from OEM vendors like CoolIT and AseTek. Aluminum tubing seems not to be used any more. Customers need to consider capital investment and running costs of liquid cooling solutions. Risk assessment of potential water leakage is another question to be aware of.

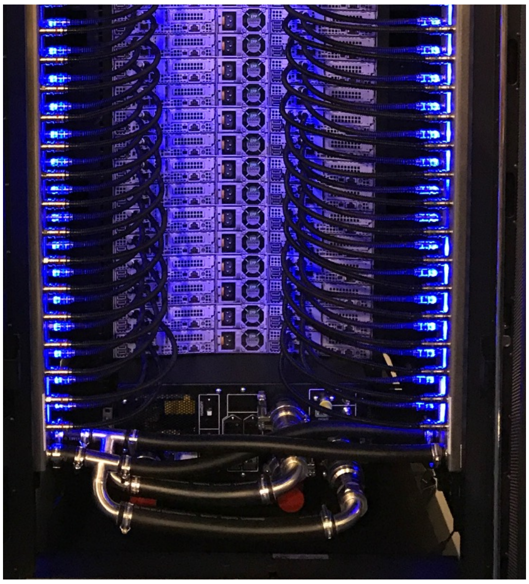

Figure 1: Mockup of a water cooling solution of the C6000 Chassis, at the show floor.

As seen in figure 1 and 2, many components of a HPC cluster does now have water cooled options.

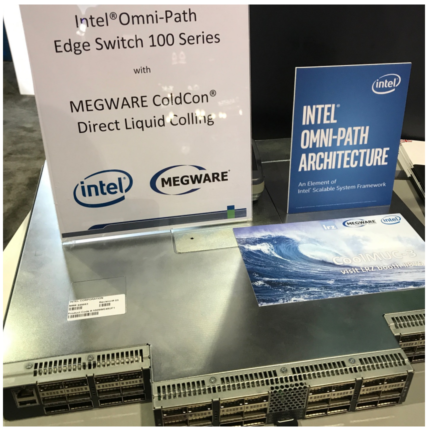

Figure 2: Water cooling on a Omni-Path switch, from MEGWARE..

Middleware / Software

Slurm

The Slurm batch queue and resource manager is open source software, and actively developed by the company SchedMD, which also offers commercial support. In a well attended BoF session a number of new features of Slurm 17.11 were described, including:

1. Federation of multiple clusters visible to all users and administrators,
2. scheduling to heterogeneous types of compute nodes.
3. X11 graphics forwarding using "srun –x11".

The future Slurm 18.08 will include support for integration with Google Cloud.

Top 500 Announcements

The 50th TOP500-list were presented at the SC17-conference. It shows that China now has 40% of the systems at the TOP500-list whereas US only has 29%. The aggregated performance per country shows that US and China are almost equal. Denmark is no longer at the TOP500-list. Norway has one system (no. 202), and Sweden has 5 systems (no. 69, 211, 253, 374 and 437).

US is far from having the most powerful computer right now, but with the two remaining computers in the CORAL-initiative coming up, USA is expected to regain the top 1-2 position(s) soon. The top-ten systems on the TOP500-list:

1. Sunway TaihuLight, China 93.0 PFlops at 15.4 MW
2. Tianhe-2, China 33.8 PFlops at 17.8 MW
3. Piz Daint, Switzerland 19.6 PFlops at 2.3 MW
4. Gyoukou, Japan 19.1 PFlops at 1.4 MW
5. Titan, USA 17.6 PFlops at 8.2 MW
6. Sequoia, USA 17.1 PFlops at 7.9 MW
7. Trinity, USA 14.1 PFlops at 3.8 MW
8. Cori, USA 14.0 PFlops at 3.9 MW
9. Oakforest-PACS, Japan 13.6 PFlops at 2.7 MW
10. K computer Japan 10.5 PFlops at 12.6 MW

Figure 3: Model of the Sunway TaihuLight cluster.

The Chinese "Sunway TaihuLight" (fig. 3) has been on the TOP500-list since June 2016. Before that, the "Tianhe-2" was no. 1 from June 2012. So it is long time since US possessed the no. 1 system. No. 4, the Japanese "Gyoukou" system, is new on the list. It has a remarkably low power consumption, which also brings it in the top of the GreenTOP500-list as number 5.

There are ongoing discussions whether the High-Performance Linpack (HPL) benchmark, which is the metric for ranking systems on the TOP500 list, still is adequate. For several reasons the High Performance Conjugate Gradients (HPCG) Benchmark has been proposed as an alternative metric. It focus on:

- sparse matrix-vector multiplication.
- vector updates.
- glocal dot-products.
- Local symmetric Gauss-Seidel smoother.

The HPCG benchmark closer reflects real life HPC applications than the HPL benchmark does. With the HPCG benchmark, the TOP500 list is shuffled quite a bit The top-five systems on the Nov. 2017 HPCG list shows[4]:

1. K computer Japan 0.603 HPCG PFlops
2. Tianhe-2, China 0.580 HPCG PFlops
3. Trinity, USA 0.546 HPCG PFlops
4. Piz Daint, Switzerland 0.486 HPCG PFlops
5. Sunway TaihuLight, China 0.481 HPCG PFlops

which brings US back in top three.

Other interesting technologies

Containers for HPC

Containers implement Operating-system-level virtualization, where Docker may be the most well-known example. However, for HPC and other multi-user systems, usage of container software gives rise to security issues. Singularity[9] is a relatively new container software developed at Lawrence Berkley National labs. It was developed with security, scientific software, and HPC systems in mind.

Delegation members participated in a Singularity workshop which discussed design principles and security issues, and participants were given a hands-on introduction to the implementation of Singularity.

Hydrogen Fuel Cells (HFC)

HPE with Daimler (Mercedes-Benz) and the National Renewable Energy Lab (NREL) presented a new project regarding Fuel-Cells with hydrogen to power green data centers. Using a car-like hydrogen engine they should be able to produce 70KW pr. unit in the data center. The system is unlimited scalable. The idea is to use this engine for the delivery of primary and backup power. The vision is to reduce energy consumption, maintenance and cost by replacing diesel power backup, UPSs, PDUs, BBUs and large cooling systems while not having to rely on grid power.

NREL proposes to produce hydrogen locally using solar cell electricity with water-splitting fuel cells with zero CO_2 emissions. However, such fuel cells are currently based upon expensive Platinum-based catalysts and have a limited efficiency. Today 95% of the hydrogen production is by steam reforming of natural gas[10], which is energy intensive and has a significant carbon footprint.

Conclusion

In conclusion, the HPC world seems to be moving faster toward exascale that most might have expected. It seems like there is an ”arms race” to get there first. There is still multiple issues to solve before the exascale mark has been reached, like cooling, power, and compute density.

In the future HPC clusters will need to be more heterogeneous, to accommodate the growing diversity in HPC jobs. As artificial intelligence becomes a bigger part of scientific compute, clusters optimized for this will be needed. As many hardware vendors are making servers based on ARM CPU now, this will likely also get traction in some HPC environments. Another trend in scientific computing is field based computations, where the compute power must be closer to the data source, e.g. analysis of trafic patterns. At SC17 there was a plenary discussing smart cities, where one of the challanges is to do the computation in real time, to provide updated information to travelers on delays, space for bikes, numbers of free seats, etc. The computation has to be done on site and cannot be sent to a data center. This opens up there will also be a bigger need for doing calculation on site, e.g. when collection data at remote location, which might not be traditional HPC, but scientific compute none the less.

Some universities in the US, have started using AI as a tool for guiding research, e.g. by analyzing which simulations is most likely to have the most interesting results.

As the exascale ”war” continues, it will be very interesting to see where the world of HPC and scientific compute goes next, and what will be the hot topic at next years SC.

Bibliography

[1] IMPROVE POWER EFFICIENCY WITH DELL EMC POWEREDGE AND COOLIT SYSTEMS RACK DCLC.
http://coolit2017.qt4egaquh7.maxcdn-edge.com/wp-content/uploads/2017/10/coolit-c6420-solutions-brief_20sept2017.pdf

[2] INTEL STORAGE Software
http://cdn.opensfs.org/wp-content/uploads/2017/06/Wed01-NeitzelBryon-LUG20Neitzel20Presentation20May20201720FINAL.pdf

[3] Lustre 2.11 and beyound
https://www.eofs.eu/_media/events/lad17/11_andreas_dilger_lad2017-lustre_2.11_and_beyond.pdf

[4] November 2017 HPCG Results
http://www.hpcg-benchmark.org/custom/index.html?lid=155&slid=293

[5] AMD Epyc 7000
http://www.amd.com/en/products/epyc-7000-series

[6] AMD Radeon mi25
https://instinct.radeon.com/en/product/mi/radeon-instinct-mi25/

[7] LENOVO Servers
https://www3.lenovo.com/dk/da/data-center/servers/c/servers

[8] BeeGFS Information
https://www.beegfs.io/content/

[9] Creating and running software containers with Singularity
https://singularity-tutorial.github.io/

[10] Hydrogen Production: Natural Gas Reforming
https://energy.gov/eere/fuelcells/hydrogen-production-natural-gas-reforming

[11] Intel drags Xeon Phi Knights Hill chips out back... two shots heard
https://www.theregister.co.uk/2017/11/16/intel_kills_xeon_phi_knights_hill/