DANISH E-INFRASTRUCTURE COOPERATION

DEIC RAPPORT 2016

FACT FINDING TOUR AT SUPER COMPUTING 16, NOV. SALT LAKE CITY, UTAH

Erik B. Madsen, Mads Boye, Nicolai Geleijns, Niels Carl Hansen, Ole Holm Nielsen, Rune Christoffer Kildetoft Andresen, Torben Kruchov Madsen.

12-12-2016

Contents

Preface

In November 2016 five Danish universities (AU, DTU, KU, SDU, AAU) participated in a DeiC funded fact finding tour. The goal was to attend the 29th ACM/IEEE Super Computing Conference (SC16), in Salt Lake City, Utah. Prior to the conference part of the delegation was invited to see Lenovo's headquarter in Raleigh, North Carolina. The delegation also attended Intel HPC Developer Conference, which was held in Salt Lake City, in the days leading up-to SC16.

During the conference, the delegation attended meetings with various hardware vendors of HPC equipment. In addition, the talks/sessions of the conference were attended based on own preferences. Please note that during the meetings with hardware vendors, information was given under Non-disclosure agreements (NDA), and therefore not all information will be disclosed in the report.

The purpose of this document, is briefly to report the delegation's findings.

CPU

AMD

After quite some years of absence from the high performance CPU market. AMD continues to be working on the server CPU "Zen" product. A detailed description was given under NDA restrictions. CPU performance is expected to become competitive with comparable Intel CPUs, with respect to performance as well as memory capacity and external connectivity. The number of cores and memory subsystem will be very attractive. Actual products are expected from multiple vendors in 2017.

AMD's return to the CPU market opens for some much needed competition on the x86_64 CPU platform. It will be interesting to see how AMD will price their new CPU's and also to see how Intel will respond to this.

Intel

The current Intel Xeon server CPU in production is the 4th generation "Broadwell". The 5th generation Xeon CPU "Skylake-SP" is anticipated to become available some time in 2017. This CPU will feature 6 memory channels at 2667 MHz, a higher core count and larger caches. There will be new networking options on-package. The details were under NDA restrictions.

Accelerators

NVIDIA

The big announcement from NVIDIA was the Pascal GPU with NVLINK. NVLINK is NVIDIAs technology fabric technology, as to not be limited by the PCI-E bus. The advantages over PCIE-E is that multiple GPU’s (up to 8) can communicate in an RDMA like way. The NVLINK fabric should provide 5 times higher performance than PCI-E. This will be available with multiple hardware vendors.

Intel

Intel is coming with a socket version of the Xeon Phi. The Phi will also have on-die Omnipath.

Flash Technologies

3D XPoint/Apache Press

Future Intel products will enable the "Apache Pass" 3D Point non-volatile RAM memory modules with very large memory sizes and running at near DRAM speeds. The Intel 3D-xPoint NVME could become a game changer with the introduction of the next Intel platform, named Purley. Apache pass along with 3D XPoint NVME will provide high-speed nonvolatile system memory and/or high speed low latency persistent storage attached to the memory bus. How this can be used is in HPC environments is still to be investigated, but it opens up the possibility to have single machines with very large memory.

Interconnect / Fabric

Mellanox

Omni-Path

Intel's OmniPath (OPA) network fabric running at 56 or 100 Gbit/s was described in detail. The OPA technology is very competitive with Infiniband, where Mellanox is the market leader. PCIe adapters and OPA switches are manufactured by Intel. Cabling using copper or fiber cables are very similar to Infiniband products. OPA installations are beginning in Europe and world-wide, and the first Danish installation is being made at DTU. The OmniPath software stack is based on OFED and is freely available. All major MPI libraries are supported. Fabric management tools are free of charge

Others

Oracle is now offering their own implementation of Infiniband as a separate product for HPC. Cables are all-optical, a clear differentiation from other vendors.

Gen-Z

A new interconnect has seen the light of day. Multiple hardware vendors showcased a new interconnect named Gen-Z. Gen-Z is a new open system interconnect which provide memory access to data and devices, either via direct-attached, switched or fabric topologies. As the interconnection is new, it will be interesting to follow how it develops over the coming years.

Servers

Meetings with DELL, HPE, Huawei was held under NDA, so there is not much which can be disclosed. In general, all vendors will upgrade their current server portfolio to support the new intel architecture.

Datacenters / server rooms

Water/Liquid Cooling

With CPU TDP power per socket going to 150W, 200W and even beyond in the near future, and GPU/accelerator power going to 300-400W per unit, the server cooling challenge has become more important than ever. Very soon traditional data center air cooling will not be sufficient for new HPC servers. When the power consumption per 42U rack increases to 40 kW, 60 kW or even 80 kW, liquid cooling inside the rack will be mandatory. The current most power-hungry servers (both CPUs and GPUs) as well as future even higher powered products will require liquid cooling inside each server. A number of liquid cooling solutions were presented at SC16, and by far the most popular product type was direct water cooling of both processors and sometimes DIMM memory modules. Other solutions included non-water coolants as well as water cooled rack doors. Water cooled data centers require the addition of heat exchanger units that convert external cooling water to the high-quality closed-loop cooling water required by servers and racks. A number of heat exchanger products were shown at SC16, among them solutions from CoolIT and the Danish company Asetek.

The ASHRAE organization publishes Standards and Guidelines for water as well as air cooled solutions. These standards will be useful for data center planning.

Rack Cooling Doors

Due to the head dissipation from CPUs, we need a way to capture and remove heat from the servers. In 'the old days', cooling units just chilled the air in the room and everything was fine - now we often have 'hot aisle' and 'cold aisle' in order to capture the heat and chill the

air.

An alternative/supplement to this, is to add active backdoors to the racks. The benefit of this design, is that it enables larger fans (at lower speed) to help move the air through the nodes. and since speed of the internal fans can be significantly reduced, the overall power consumption can be reduced to. However, this design requires tubes with water to every rack, so it needs to be part of the planning or redesign of a site - and it still relies on air to move the heat from the CPU.

Every CPU has a TDP (Thermal Design Power) and in the context of cooling, it's roughly the amount of heat we need to remove, in order to keep the CPU running (at maximum performance) - and thus dictates how large the heat sink on the CPU, needs to be. With the current generation of

CPUs, it is possible to fit 2 nodes for each unit of rack space - but for the next generation of CPUs, we are looking at a significant increase in TDP - so it will be an challenge to fit 2 nodes (with a sufficiently large heat sink) within a unit of rack space. This can be solved by either accepting to only have 1 node pr. unit of rack space or use something better than air to cool the CPUs.

At SC'16, we saw a lot of solutions with liquids (often water) - some manufacturers integrated liquid cooling in their design, while others referred to 3. party companies which specialize in liquid cooling. As all of them requires water to the node, this will be a key feature to consider, when designing a new site or upgrading an existing.

Middelware / Software

Slurm

Slurm as a batch queueing and resource manager Open Source software is gaining a strong usage across the world. Slurm is extremely rich in features, and is actively developed due to requirements of customers. Features of the latest and future versions of Slurm were described at the Slurm booth as well as at a BOF sessions. The BOF session attracted 200+ people and had a lively discussion. Slurm is used on many large HPC centers in USA and Europe, and at a large number of regional and local HPC centers. The quality of the software, the Open Source license, as well as the option of paid support makes Slurm an attractive solution.

Intel Compiler

Intel has put a lot of work into optimizing Intel Python, which was announced last year. They have made a collaboration with Continuum, which makes the Anaconda python distribution, to optimize Intel python. A lot of work has gone into optimizing towards MKL libraries, and the Numba Jit compiler for python

OpenHPC

The openHPC BOF did not give a clear picture of where the project is going. There is a lot of good initiative in the project, but it is still very new, and have not found a solid strategy in terms of release cycle, development focus. The project still seems very new, it was founded at SC15, but a lot of good work has gone into it, and definitely worth keeping an eye on in the time to come.

HPC Orchestrator

HPC Orchestrator is a new product from Intel, which is a fork of the OpenHPC project. This will allow intel to focus the on better support and integration with both Xeon Phi and in Intel CPUs.

Top 500 Announcements

Selected entries from the 48th TOP500 list, November 2016:

Rank	Name	Country	Arch	Performance
1	Sunway TaihuLight	China	Sunway, 1,45 GHz	93 Pflops
2	Tianhe-2	China	NUDT/Intel, 2.2 GHz + Xeon Phi	33.8 Pflops
3	Titan	ORNL, USA	Cray XK7/AMD 2.2 GHz + Nvidia K20x	17.6 Pflops
8	Piz Daint	CSCS, Swiss	Cray XC50/Intel 2.6 GHz + Nvidia P100	9.7 Pflops
425	Computerome	DTU, Denmark	Hpe/Intel 2.0 GHz	410 Tflops
500	Platform	USA	Hpe/Intel 2.4 GHz	349 Tflops

USA and China both have 171 systems on the list, but China has the two systems in the top. Wrt. *aggregate* Linpack performance USA leads by 33.9% of the total performance of the systems on the TOP500 list, closely followed by China with 33.3%. Germany, Japan and France follows with 31, 27 and 20 systems respective.

The total performance of all 500 computers on the list is now 672 petaflops, a 60% increase from a year ago, which corresponds to the expected exponential growth.

Also the growth in performance of #500 on the list is about 60% compared to the year before. The progress in performance is unchanged over more than two decades, which for some is the proof that Moore's law still applies.

The number 1 system comprises not less than 10.6 million Sunway-cores manufactured in China. The performance of this huge system is equal to the next 5 systems on the list and draws 15.3 MW!

The number 1 system in Europa is still the “Piz Daint” system in Switzerland. Since the last list, it has been upgraded with new Nvidia P100 GPUs.

The number of systems with accelerators (f.ex. Nvidia GPUs and Intel Phi) has dropped from 103 in 2015 to 86 on this year list.

Computerome at DTU is now number 425, a year ago it was number 236.

Abacus2.0, the number 267 last year, is not even in top 500 anymore (NB: if a new benchmark had been carried out after Abacus2.0 was upgraded with more nodes this spring, it probably still would have made it on the list). Number 500 of this year, “Platform” doing 349 Tflops, would have been ranked as number 404 in last year list.

Super Computing 17

The next SC conference will be held in Denver, Colorado in November 20016.