
SC15 Austin, TX | hpc transforms.

# SC15 report

Andrei Gheata

SFT group meeting, November 23, 2015

# Venue

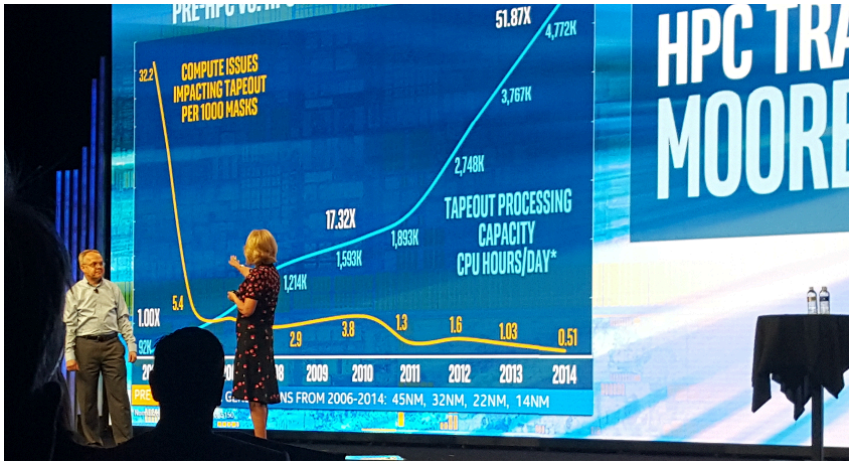

- 28-th edition, SC is a HPC conference held yearly in US
  - Among the biggest HPC conferences
- Austin, TX, 15-20 Nov 2015
  - Austin Convention Center
  - Participants: 10000+ people, 352 exhibitors (industry/ research)
- SC16: Salt Lake City, UT 13-18 Nov 2016

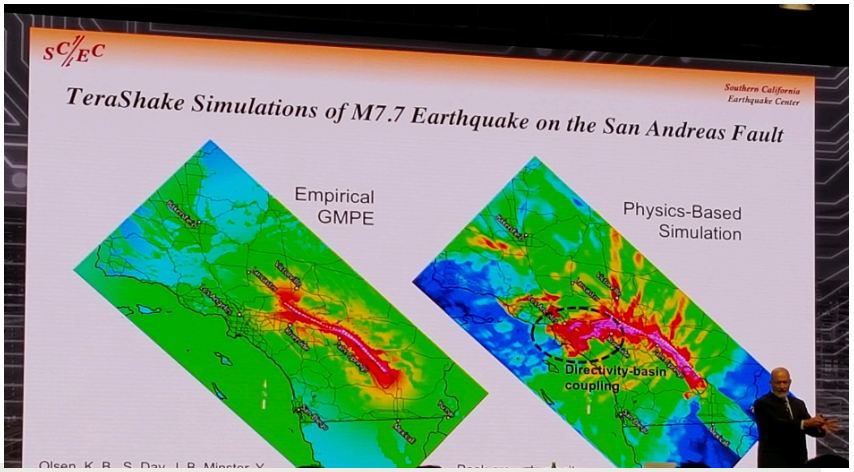

# Program (or how to get HPC-intoxicated…)

- Technical program
  - Plenary/invited talks
  - Technical papers/posters
  - Tutorials
  - Panels (get the word from the guru's)
  - Workshops
  - Scientific Visualization Showcase
  - Birds of a Feather (audience driven discussions)
  - Awards
- Exhibits
  - Latest technologies and discoveries from industry and research
- Students@SC
  - Cluster competition, student volunteers & experiencing HPC 4 undergraduates
  - Mentor-Protégé Program, Student-Postdoc Job Opportunities

# Keynote talks

- HPC Matters plenary session given by Diane Bryant
  - Intel datacenter business unit leader, speaking on the importance and directions of HPC
  - Simulation driven science -> Data model driven science
    - Data analytics/machine learning
  - "technology goes 5nm…" – evolution or revolution?
  - Exascale limited by: power efficiency/ cost per performance/accesibility to all
  - https://youtu.be/kuh5qzZI2HM
- Several invited talks on many subjects



# Alan Alda – an inspiring intro

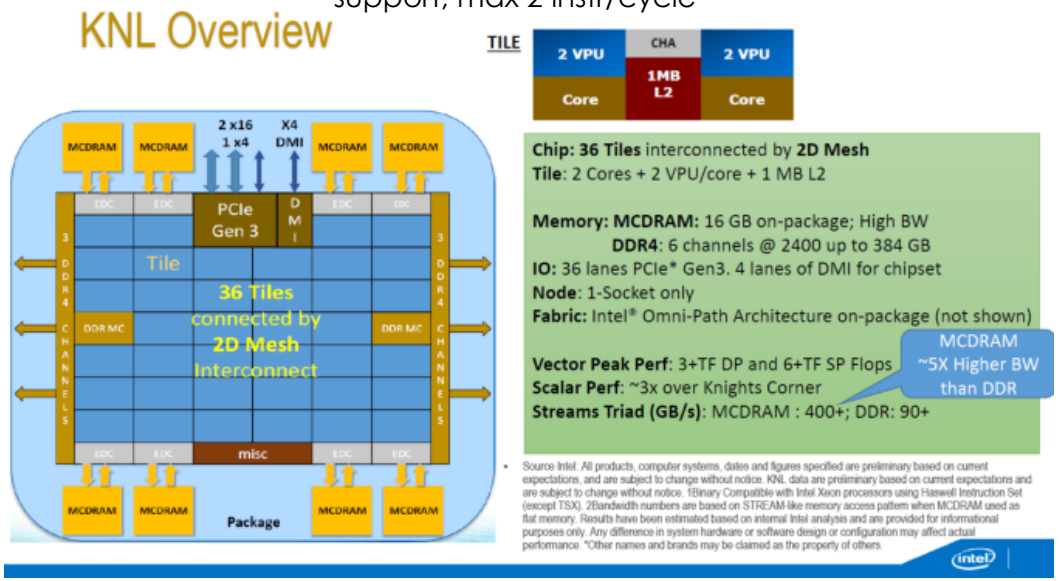
- Actor, writer, science advocate, and Visiting Professor at Stony Brook University
  - Shared his passion for science communication and its importance
- Analogy with 3 phases of love
  - Attraction: body language and tone of voice prevail on words
  - Infatuation: think about all time, memory is helped by emotion
    - "put a little word of emotion…"
  - Commitment: listen to each other, empathy
- "Curse of knowledge" – others may not know what you know…
  - Try to improvise, tell a (dramatic) story!
  - Do not explain jargon with more jargon
  - Try to explain to an 11 years old


*M*A*S*H (1972-1983)*

# Intel MTA: the Knights Landing

- 2nd generation Phi
  - Bootable processor
  - PCIe & Omni-Path versions later
  - **3+ TF, 3x KNC** single thread performance
  - **16 GB MCDRAM@450 GB/s + 90 GB DDRAM ➔ high BW workflows**
- Chip on tiles design
  - 36 tiles (2 cores) 2D mesh interconnect
  - Dynamic partition of resources to threads
    - **1 thread can saturate**
- 32 vector reg, & 8 mask reg.
  - Gather/scatter
- Monitor/Mwait instruction
  - Set HW points/resume rather than spin

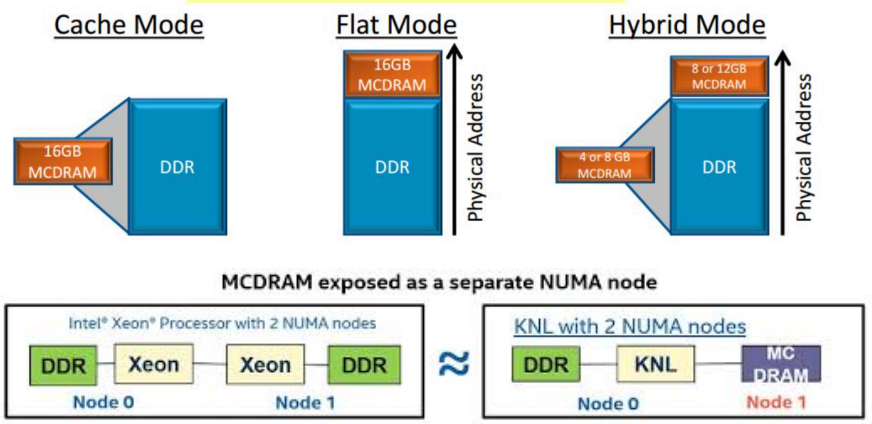
4 thr/core, fast unaligned access, gather/scatter support, max 2 instr/cycle



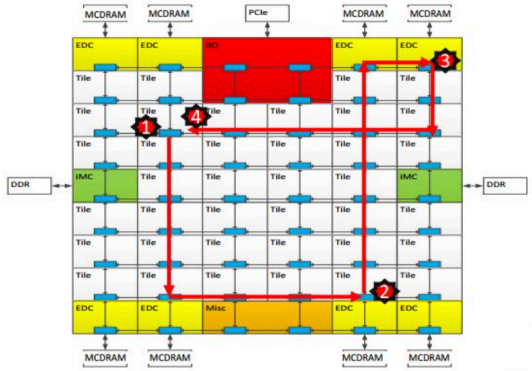

- AVX-512 extensions
  - Conflict Detection (vect. portions of loop)
  - Prefetch instructions (gather/scatter)
  - Exponential and reciprocal

# KNL – caching & locality

- Several **memory modes** selectable at boot
  - Cache mode: MCDRAM is a cache of DDR
    - to avoid if many cache misses
  - Flat mode: MCDRAM (node 0) & DDRAM (node 1) sharing the address space
    - High bandwidth guaranteed, but SW has to manage NUMA allocations & copy
  - Hybrid mode: 4-8 GB MCDRAM cache mode, 4-8 shared
- KNL mesh interconnect
  - Passing cache lines across tiles, via distributed directories, to memory
  - All-to-All – no affinity – largest latency
  - Quadrant – affinity directory-memory (transparent to SW)
  - SNC (Sub-NUMA-Clustering) ⇔ quad core Xeon




Cluster Mode: All-to-All



# Intel MTA: 3D XPOINT – NVM solutions

- NVM express protocol – interface PCIe <-> SSD
  - Ready for Intel's Optane brand of NVM
    - **1000x faster** than flash
- 2 trends (2016/2017)
  - 3D NAND 48/32 GB – block architecture
  - 3D Xpoint with **~0 cost IOPS**
    - **10x density of RAM**
    - 500K writes (4K blocks)
    - <10 us latency at 99%, <60 us at 99.999% quality of service
- New paradigm in cluster computing
  - Hot gets hotter
- ColdStream 3D Xpoint SSD's
  - Up to 3TB
  - 550K IOPS R/W
  - 10 us latency
  - 2500 MB/s read/write rate
  - Active/idle power: 18W/3W
  - "Wicked" fast – can saturate the RAM
- Elkdale dual port NVRAM D3X00 – Q1 2016
- Coldstream Dual Port – Q1 2017
  - 375 GB-> 1.5 TB
- Intel SSD DC P4500 -> "Cliffdale" ->8TB

# Intel MTA: Apache Pass DIMM

- Server memory architecture based on 3D XPoint tech.
  - Plugs on standard DDR4
  - New class of memory
  - 1-RAM, 2-DISK, **3-AppDirect**
- "**Jumbo**" memory
  - Virtualization, big data and cloud in memory DB
- Memory resilience
  - Persistence to power cycles
- Hyper-speed storage
  - Use all IOPS you can get
- Performance gap SSD <-> RAM still huge, filled by 3D Xpoint
  - ApachePass+Coldstream
  - 6TB memory (3TB/CPU), behaving like memory (volatile)
  - One can carve out and emulate the disk in this space
- AppDirect mode – partition modes (memory/disk) on the fly
  - **Carve out piece and give it to the app**
  - AES256 encryption

# Workshops, posters, papers

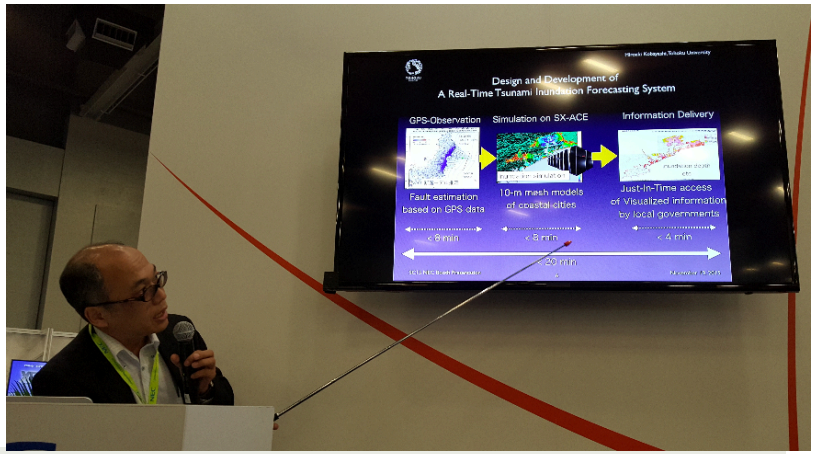
- 22 full day, 20 half day workshops, targeting subjects and communities in scientific/software communities
  - Sunday, Monday and Friday (minimize overlap with exhibits)
  - Both ad-hoc on specific (submitted) subjects or regular workshops
  - High probability to find an interesting subject
  - Not archived in the proceedings as of SC13
- ~140 posters on high performance computing, storage, networking and analysis
  - Archived digitally and made available after the conference
  - ACM Student Research Competition posters
  - Awards for posters
- ~85 papers, written BEFORE the conference and made available on a CD

# Awards – an example

- "How to teach exascale machine to do the data dance"
  - Ken Kenedy award – Katherine Yelick (LBNL) – DOE program coordinator
- Exascale = order of magnitude increase in performance at all scales, not just exaflop
- Image analysis, data extraction in "streaming" datasets
  - Palomar transient factory (systematic exploration of the dynamic sky), biology (Gene Context Analysis)
- Random access analytics – genome assembly
  - Random access in big memory -> huge hash tables, needing low overhead communication
- Data productivity
  - Spark – 100x faster than Hadoop MapReduce in memory, 10x on disk
  - Cloud computing

# Exascale: NEC SX-ACE

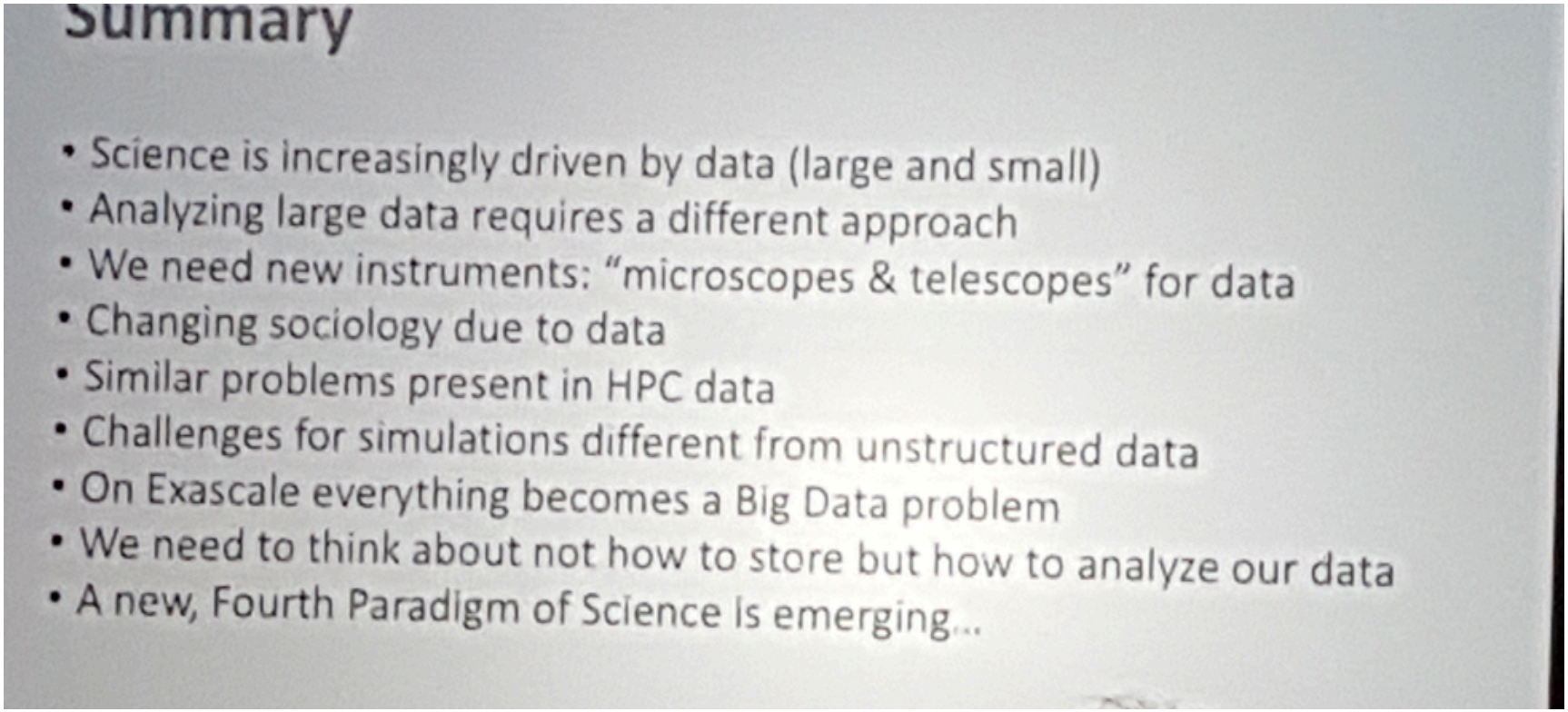
- SX-ACE = new generation NEC Vector Supercomputer
  - Designed for memory bandwidth intensive applications
  - SX2/3 (1983) -> SX8 -> SX-ACE
    - 64 GFlops core, 3GB/s/Watt
- AURORA project – best bandwith/$
  - Several societal projects in AURORA vision to use SX-ACE
  - Tsunami real-time simulation system
    - 20 min response time from earthquake rupture for detailed flooding simulation
    - Detailed imaging in the government office
    - Probes detecting the seismic wave, feeding simulation engine

# Panels

- Many HPC subjects
  - "Post Moore's law computing: Digital versus Neuromorphic versus Quantum"
  - "Future of Memory Technology for Exascale and Beyond"
  - "Supercomputing and Big Data: From Collision to Convergence"
  - "Programming Models for Parallel Architectures and Requirements for Pre-Exascale"
  - Asynchronous Many--Task Programming Models for Next Generation Platforms
  - …
- Some lead to interesting technical discussions
- Many diverged into endless philosophical discussions (or dissertations on personal views on HPC)
- Many questions but not necessary as many answers…

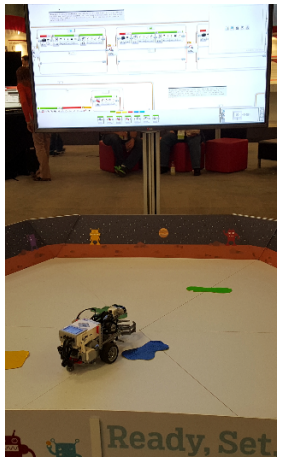
# Panels


Summary
• Science is increasingly driven by data (large and small)
• Analyzing large data requires a different approach
• We need new instruments: "microscopes & telescopes" for data
• Changing sociology due to data
• Similar problems present in HPC data
• Challenges for simulations different from unstructured data
• On Exascale everything becomes a Big Data problem
• We need to think about not how to store but how to analyze our data
• A new, Fourth Paradigm of Science is emerging...

# Exhibitions – the show-off

# Anything from gadgets to exascale computing solutions




# Intel tours

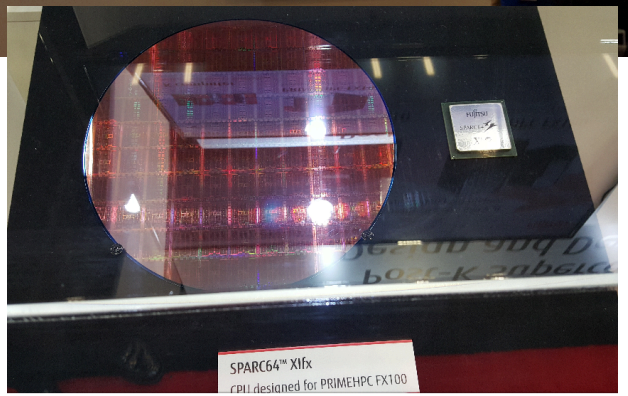

FUEL YOUR
INSIGHT
NetApp
FUJITSU


Storage & Blade
Solutions
Twin Architecture
Omni-Path
intel


CRAY


FUJITSU Supercomputer
PRIMEHPC FX100
MPI Collective Communication Performance
ARISTA


intel
XEON PHI
inside

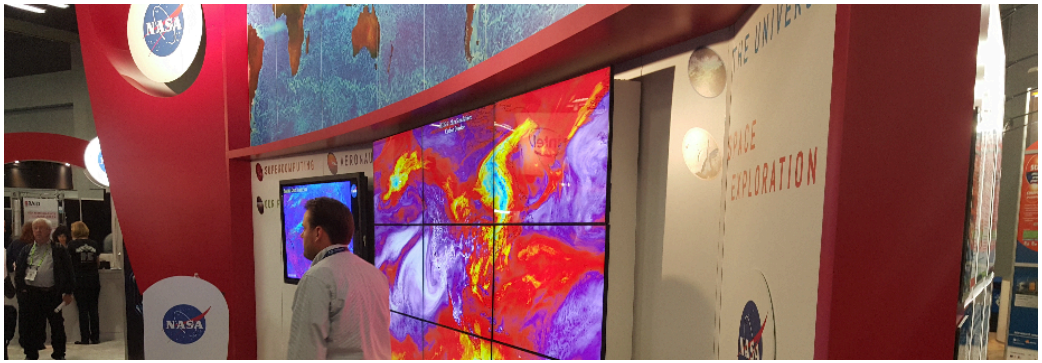

SPARC64™ XIfx
CPU designed for PRIMEHPC FX100

# Networking, memory, storage

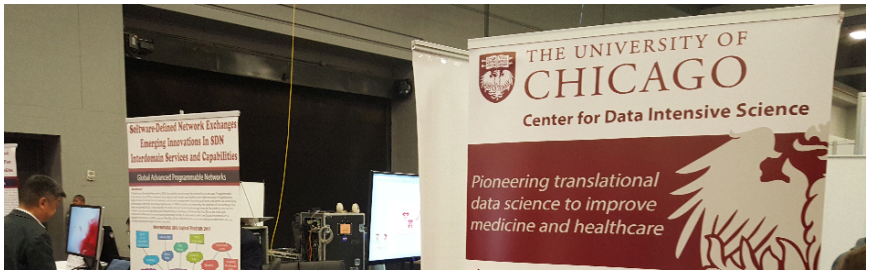




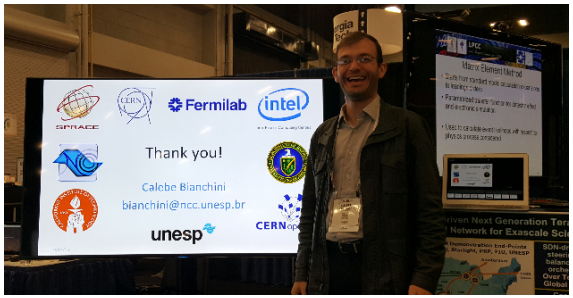
# Labs, universities, computing centers

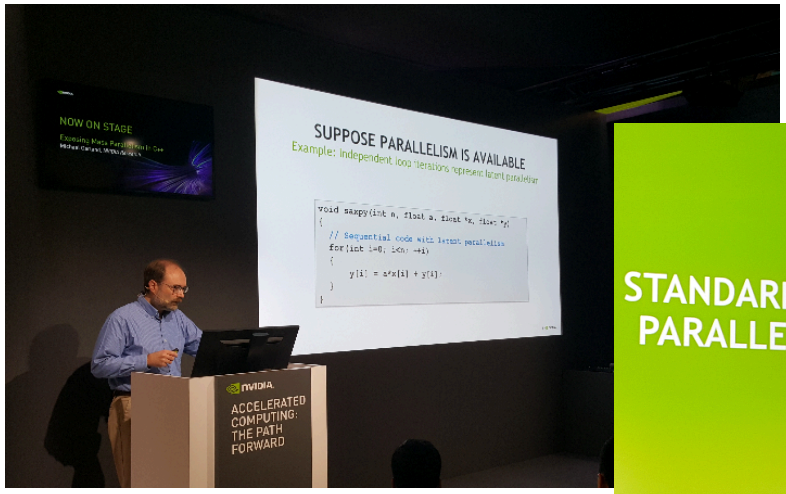


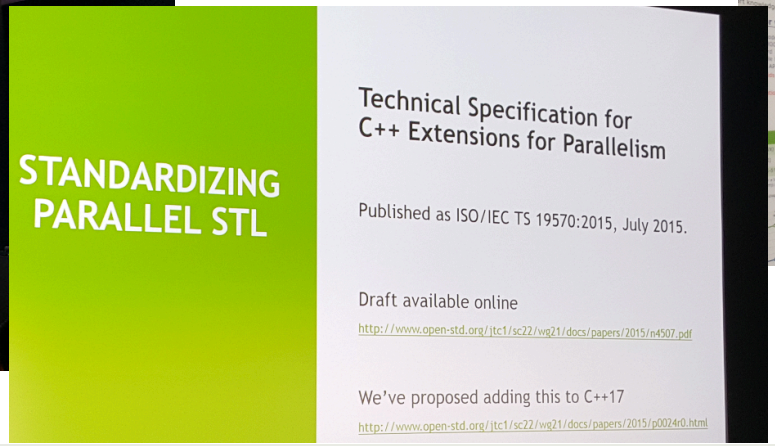


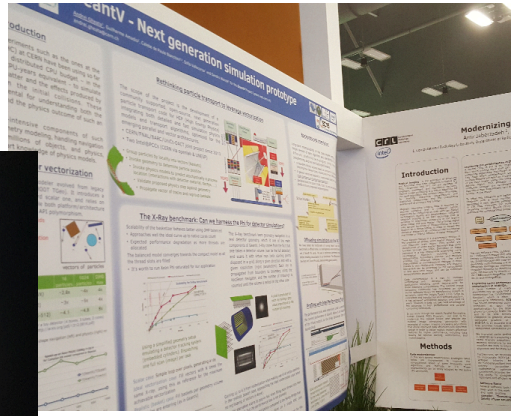



# Hands-on, invited talks on booths, posters

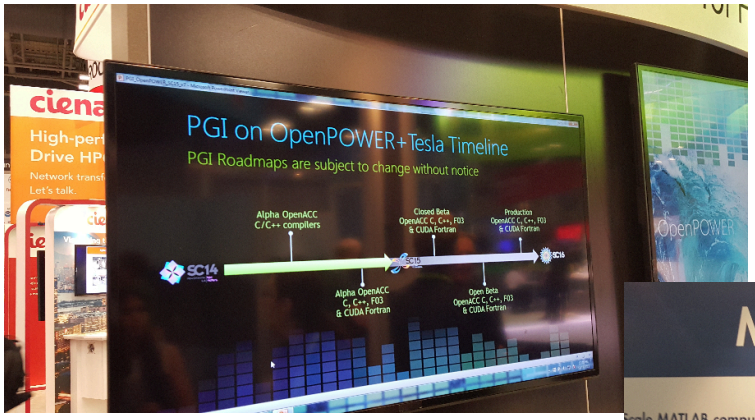



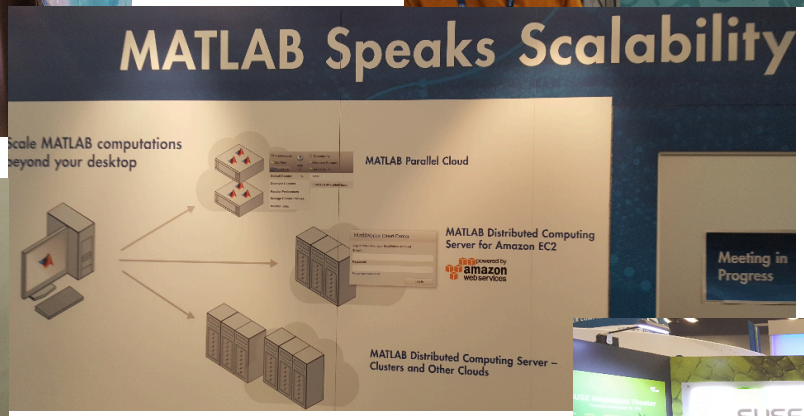



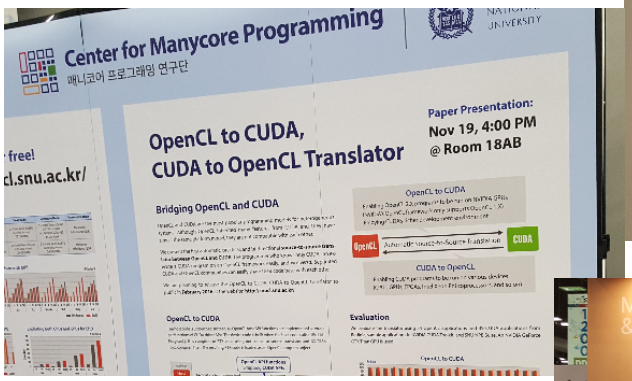



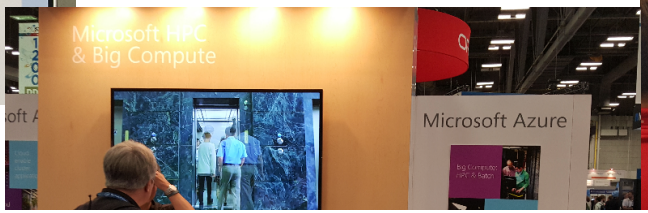
# Software, compilers








# My impressions

- More than a conference, a big gathering mixing scientific computing and technology
  - "Super" goes well with the name, has everything may cross your mind
  - One has to target what he/she is going for, cannot cover it all
  - Big show with exhibits, everybody trying hard to make a scientific/ technical showcase
  - Important occasion for discussing business with partners/providers
- Data analytics (big data, machine learning, …) getting most focus
  - Technology is what allowed this to happen
    - High-throughput computing pushed from every side: memory BW and low latency, high density storage more FLOPS/Watt, faster fabric interconnects
  - More and more HW features becoming SW programmable
    - Does not make our life easier…
- There's something for everyone