Argonne NATIONAL LABORATORY

The Coming Convergence of HPC and Big Data

Pete Beckman

Senior Scientist, Argonne National Laboratory

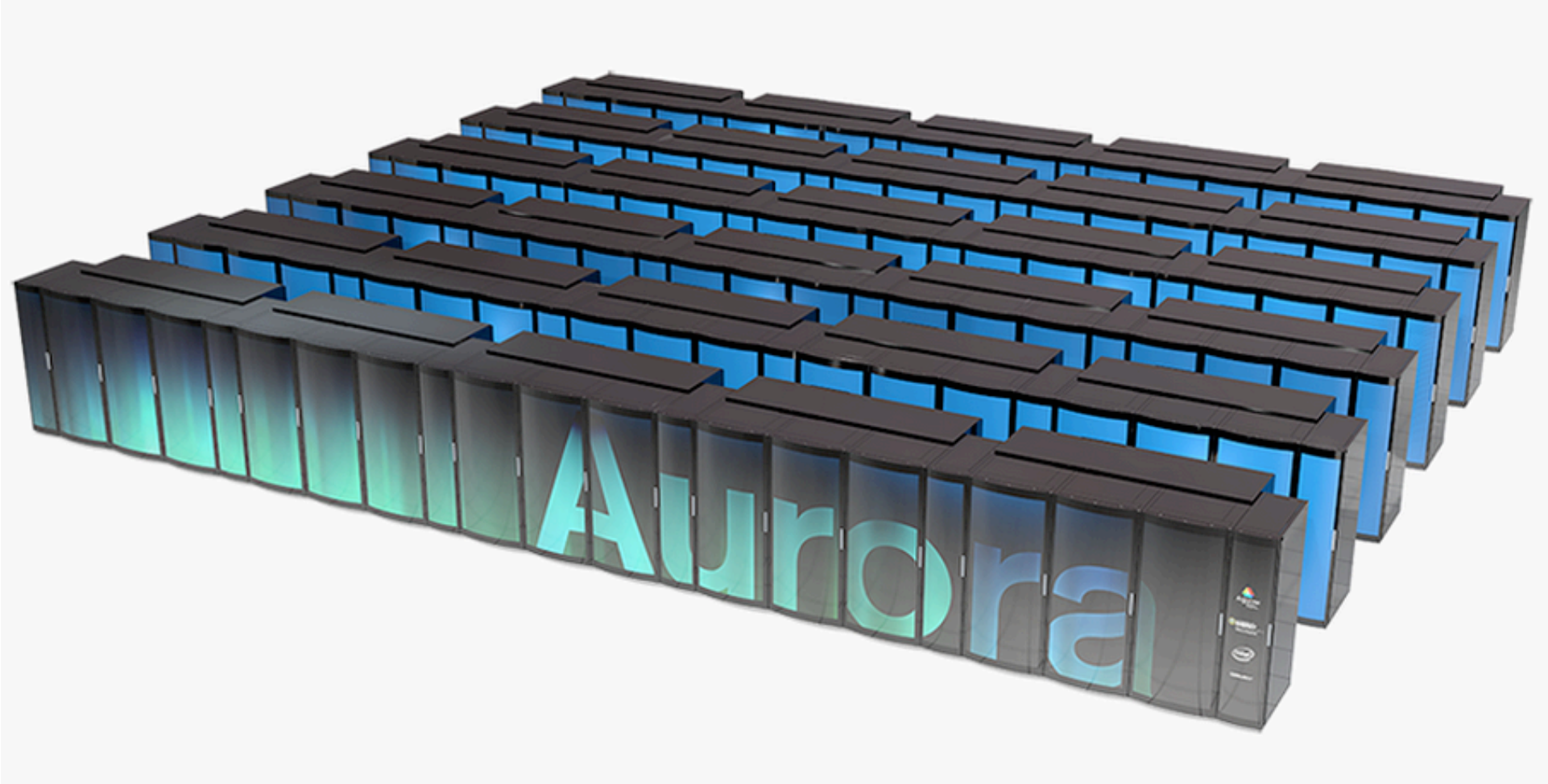
Co-Director, Northwestern / Argonne Institute for Science and Engineering (NAISE)

Senior Fellow, University of Chicago Computation Institute

U.S. DEPARTMENT OF ENERGY


Argonne National Laboratory

Argonne's Next Big Machine: Aurora



Argonne's Aurora Details

System Feature	Aurora
Peak System performance (FLOPs)	180 - 450 PetaFLOPS
Processor	3rd Generation Intel® Xeon Phi™ processor (code name Knights Hill)
Number of Nodes	>50,000
Compute Platform	Cray Shasta next generation supercomputing platform
High Bandwidth On-Package Memory, Local Memory, and Persistent Memory	>7 PetaBytes
System Interconnect	2nd Generation Intel® Omni-Path Architecture with silicon photonics
Interconnect interface	Integrated
~~Burst Storage Buffer~~ NVRAM	Intel® SSDs, 2nd Generation Intel® Omni-Path Architecture
File System	Intel Lustre* File System
File System Capacity	>150 PetaBytes
File System Throughput	>1 TeraByte/s
Intel Architecture (x86-64) Compatibility	Yes
Peak Power Consumption	13 Megawatts
FLOPS/watt	>13 GFLOPS/watt
Delivery Timeline	2018
Facility Area	~3,000 sq. ft.

Europe-USA-Asia Workshops on Big Data and Extreme Computing

Following the International Exascale Software Initiative
(IESP 2008-2012 ➔**Big Data and Extreme Computing** workshops (BDEC)
http://www.exascale.org/bdec/

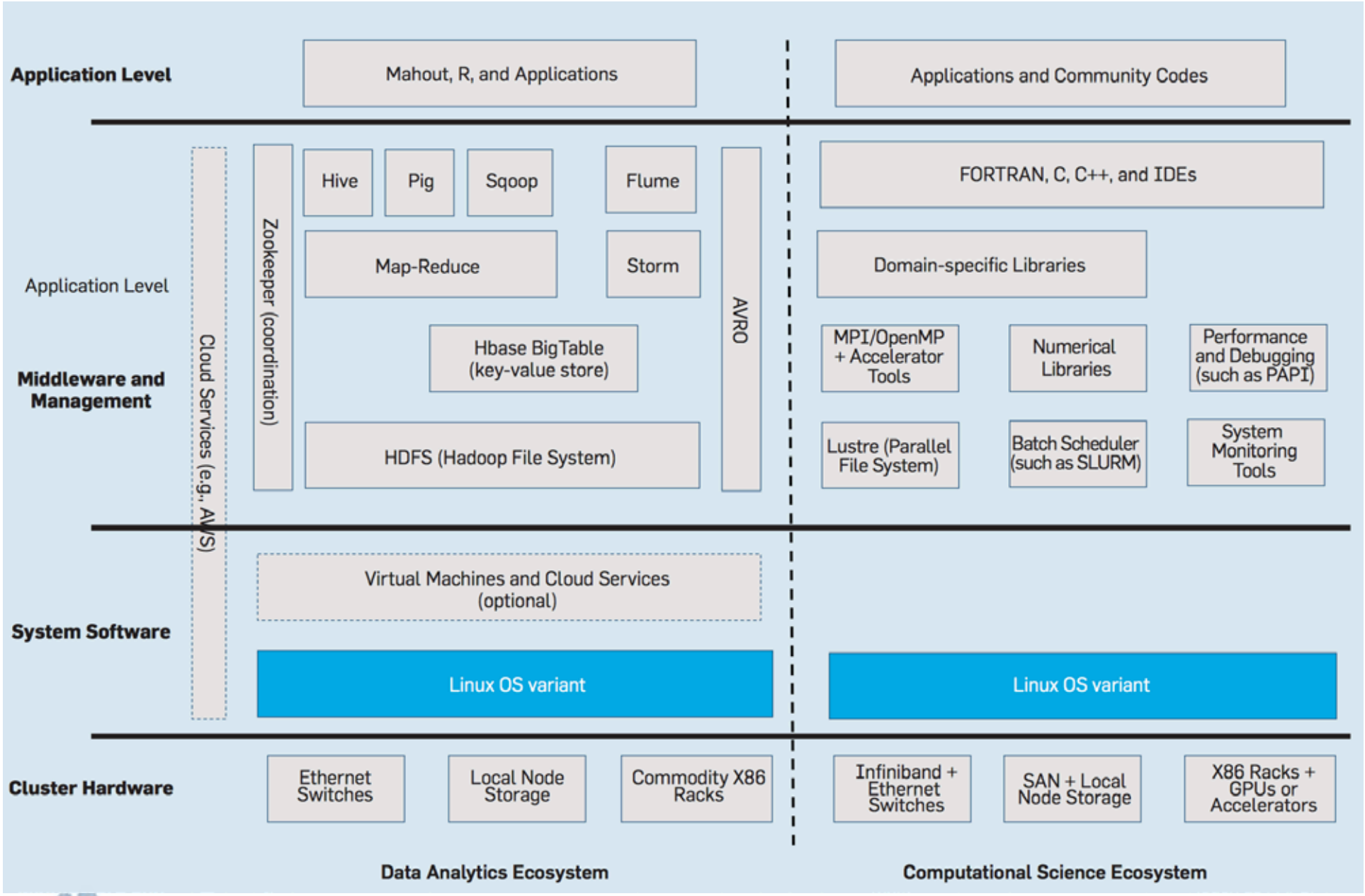
Overarching goal:

1. Create an international collaborative process focused on the co-design of software infrastructure for extreme scale science, addressing the challenges of both extreme scale computing and big data, and supporting a broad spectrum of major research domains,
2. Describe funding structures and strategies of public bodies with Exascale R&D goals worldwide
3. Establishing and maintaining a global network of expertise and funding bodies in the area of Exascale computing

1 – BDEC Workshop, Charleston, SC, USA, April 29-May1, 2013

2 – BDEC Workshop, Fukuoka, Japan, February 26-28, 2014

3 – BDEC Workshop, Barcelona, Spain, January 28-30, 2015


Application Level
Mahout, R, and Applications
Applications and Community Codes
Application Level
Middleware and Management
System Software
Cluster Hardware
Cloud Services (e.g., AWS)
Zookeeper (coordination)
Hive
Pig
Sqoop
Flume
Map-Reduce
Storm
AVRO
Hbase BigTable (key-value store)
HDFS (Hadoop File System)
FORTRAN, C, C++, and IDEs
Domain-specific Libraries
MPI/OpenMP + Accelerator Tools
Numerical Libraries
Performance and Debugging (such as PAPI)
Lustre (Parallel File System)
Batch Scheduler (such as SLURM)
System Monitoring Tools
Virtual Machines and Cloud Services (optional)
Linux OS variant
Linux OS variant
Ethernet Switches
Local Node Storage
Commodity X86 Racks
Infiniband + Ethernet Switches
SAN + Local Node Storage
X86 Racks + GPUs or Accelerators
Data Analytics Ecosystem
Computational Science Ecosystem

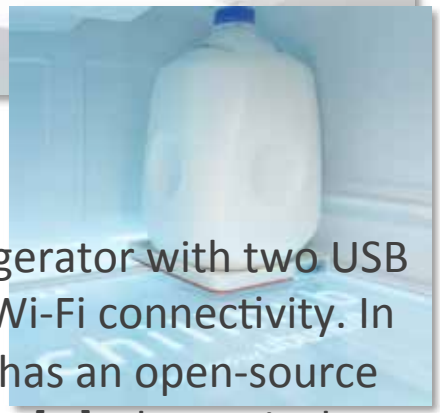
Internet of Things… Sometimes Silly

University of Cambridge, 1991

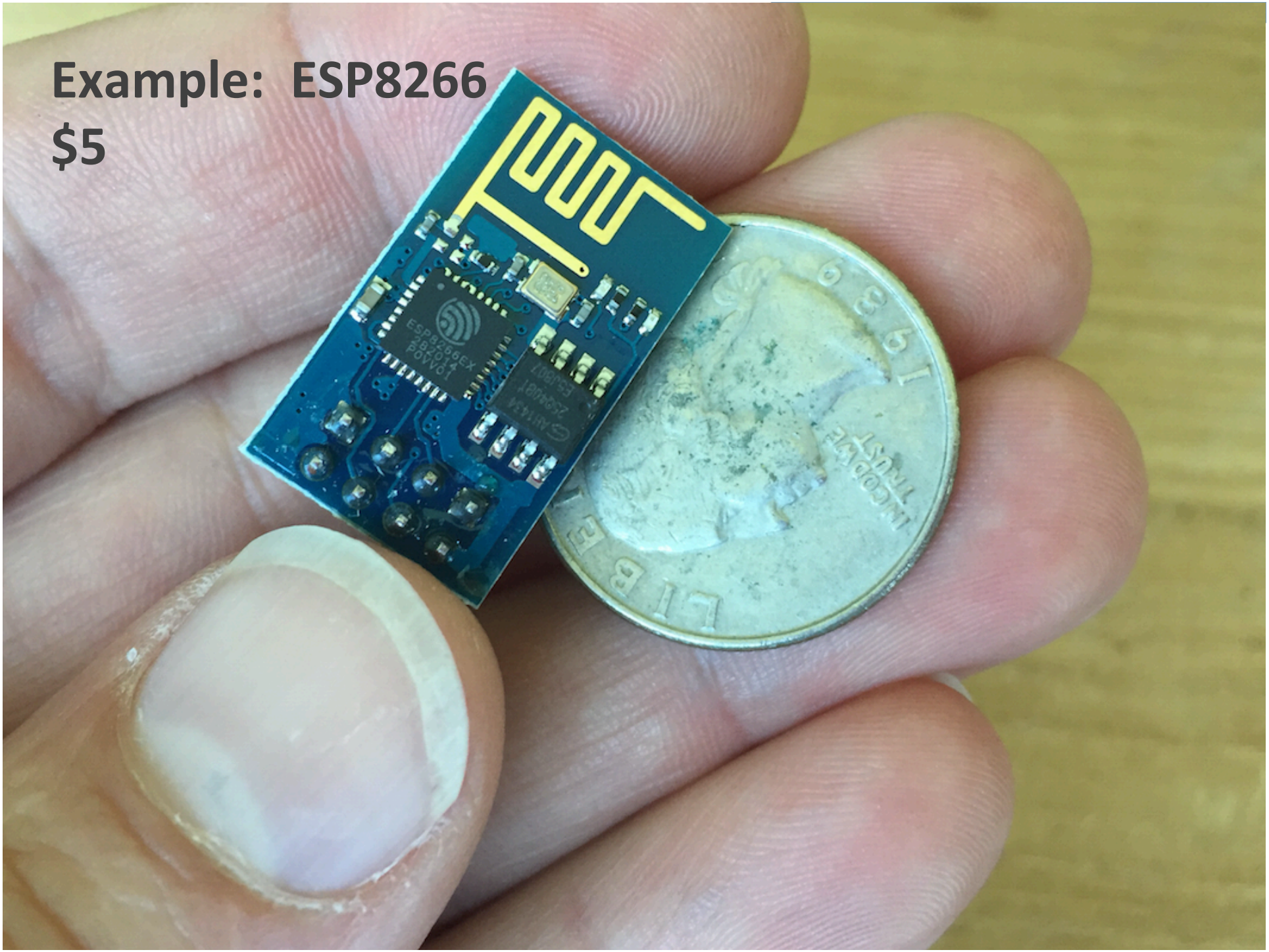
Amazon Dash

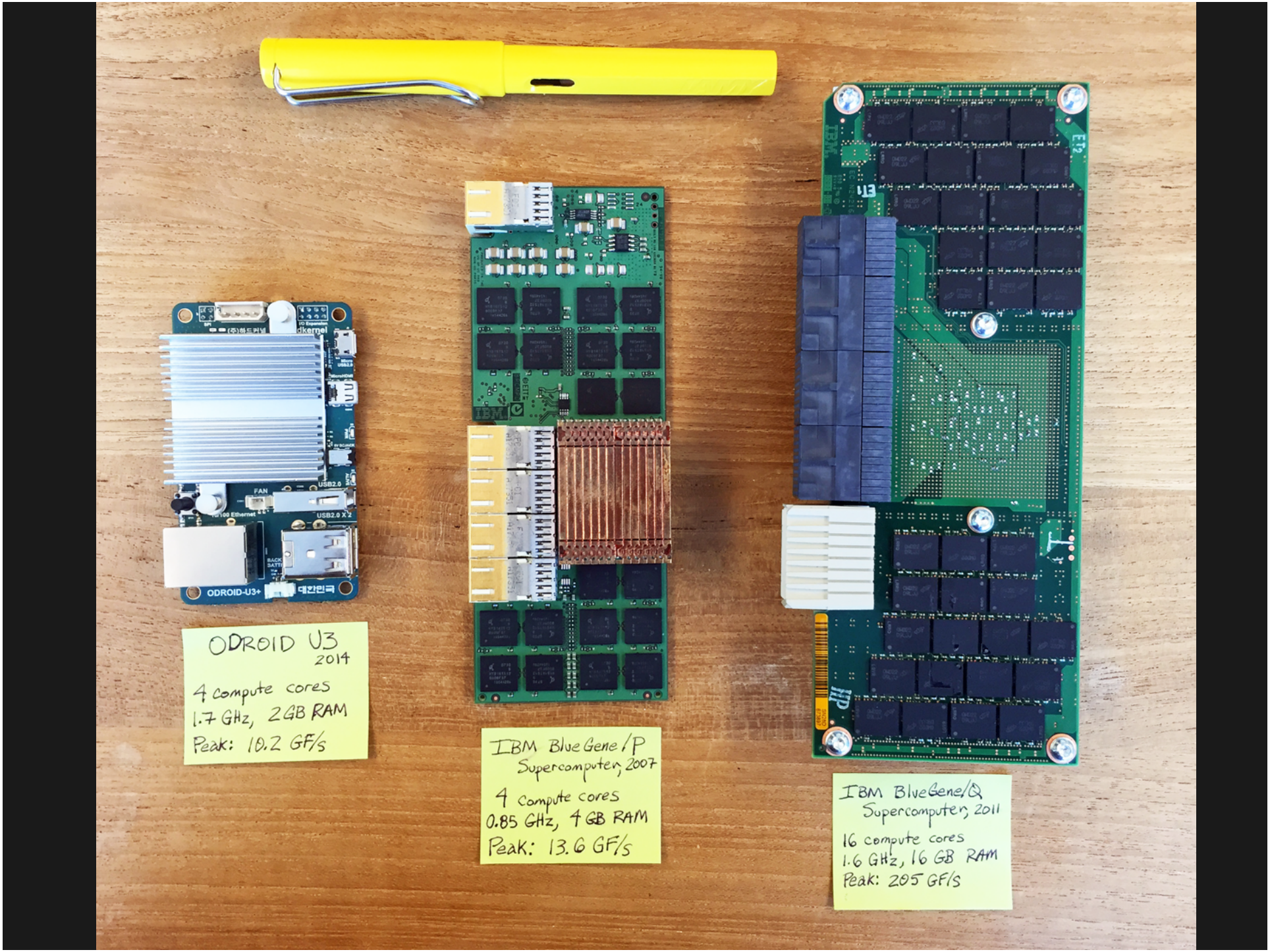
IoT: Canonical and GE's FirstBuild Collaborate on Smart Refrigerator

By Canonical on 11 May 2015



"ChillHub is a refrigerator with two USB ports and built-in Wi-Fi connectivity. In addition, ChillHub has an open-source iOS-compatible app […] Ubuntu is the favored platform for developers of all kinds – particularly those innovating around the Internet of Things."

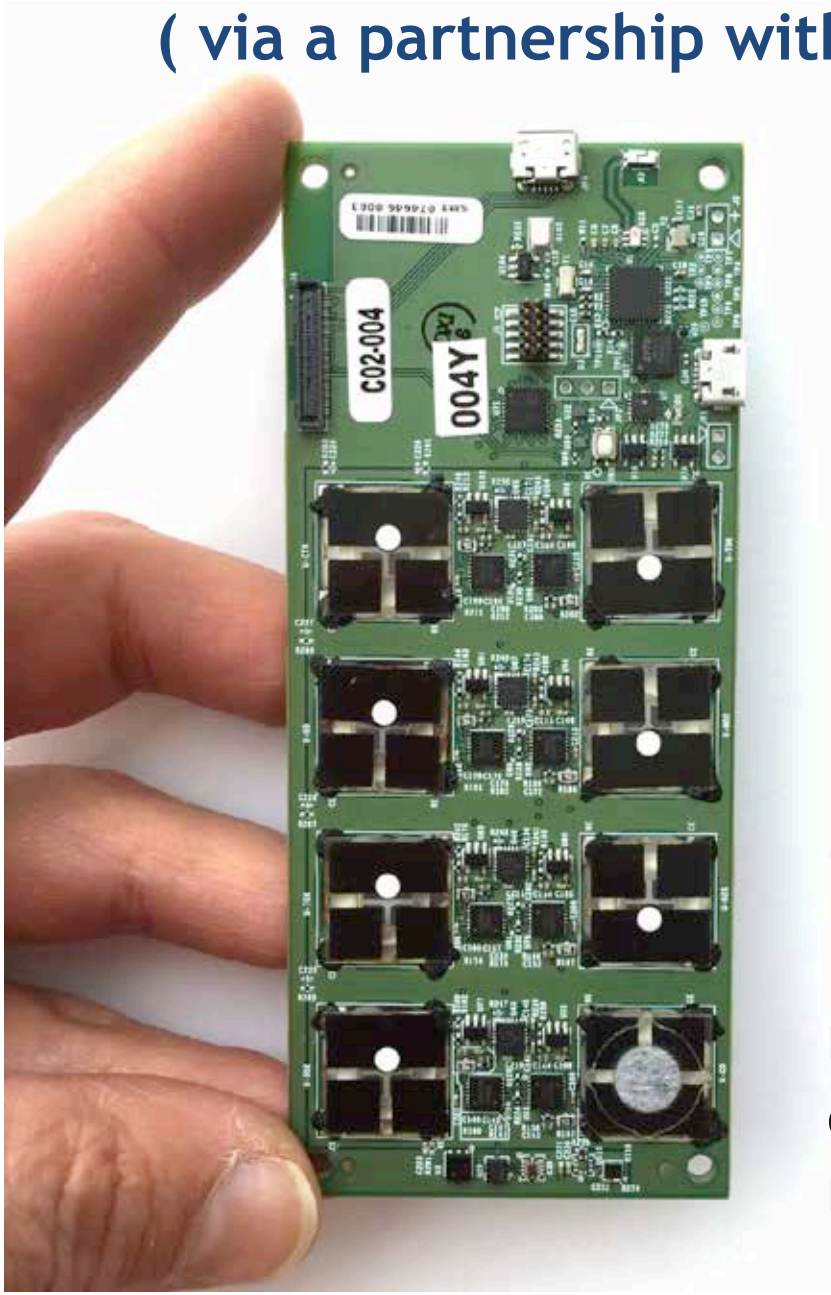

Example: ESP8266
$5


ODROID U3
2014
4 compute cores
1.7 GHz, 2GB RAM
Peak: 10.2 GF/s
IBM BlueGene/P
Supercomputer, 2007
4 compute cores
0.85 GHz, 4 GB RAM
Peak: 13.6 GF/s
IBM BlueGene/Q
Supercomputer, 2011
16 compute cores
1.6 GHz, 16 GB RAM
Peak: 205 GF/s

An open software and hardware, hackable, wireless sensor platform with in-situ computing.

http://www.wa8.gl

New Advanced Sensors
(via a partnership with Intel)



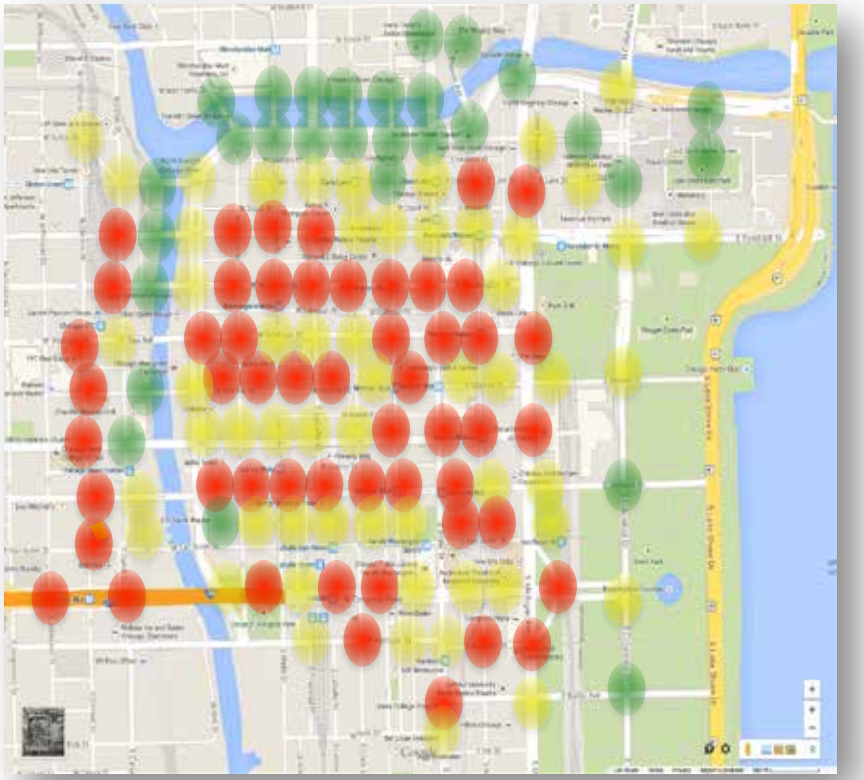
- NO_2 (Nitrogen Dioxide): <2 ppb
- O3 (Ozone) < 5 ppb
- CO (Carbon Monoxide) < 1 ppm
- SO2 (Sulfer Dioxide) < 15 ppb
- H2S (Hydrogen Sulfide) < 2 ppb
- TOX (total oxidizing index) < 1 ppm CO equiv
- TOR (total reducing index) < 2 ppb NO2 equiv
- Future:
  - HCHO (Formaldehyde)
  - VOC (Volatile Organic Compound)
  - CH4 (Methane)

Intel is partnering with SPEC Sensors to distribute a production version of the device. Pricing TBD.


GERALD

Artist Rendition

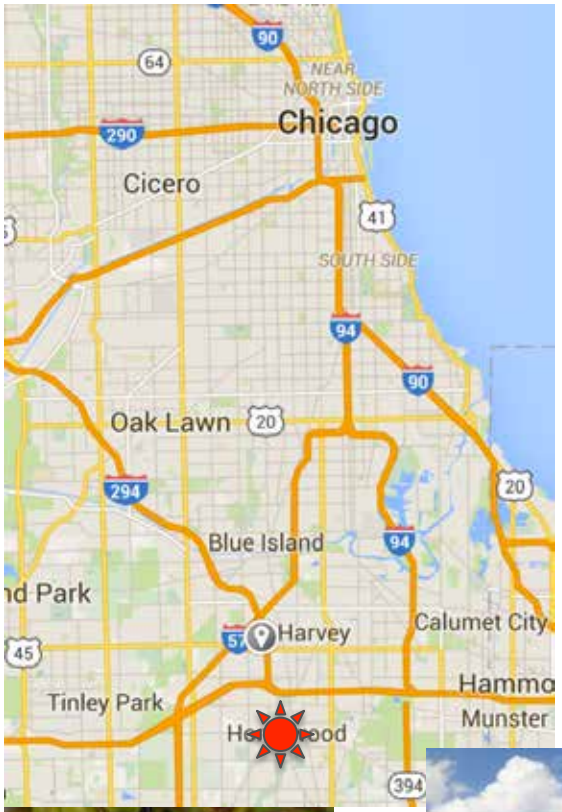
ARRAY OF THINGS

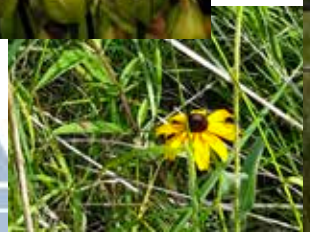


Deployment Map

Gensburg-Markham Prairie

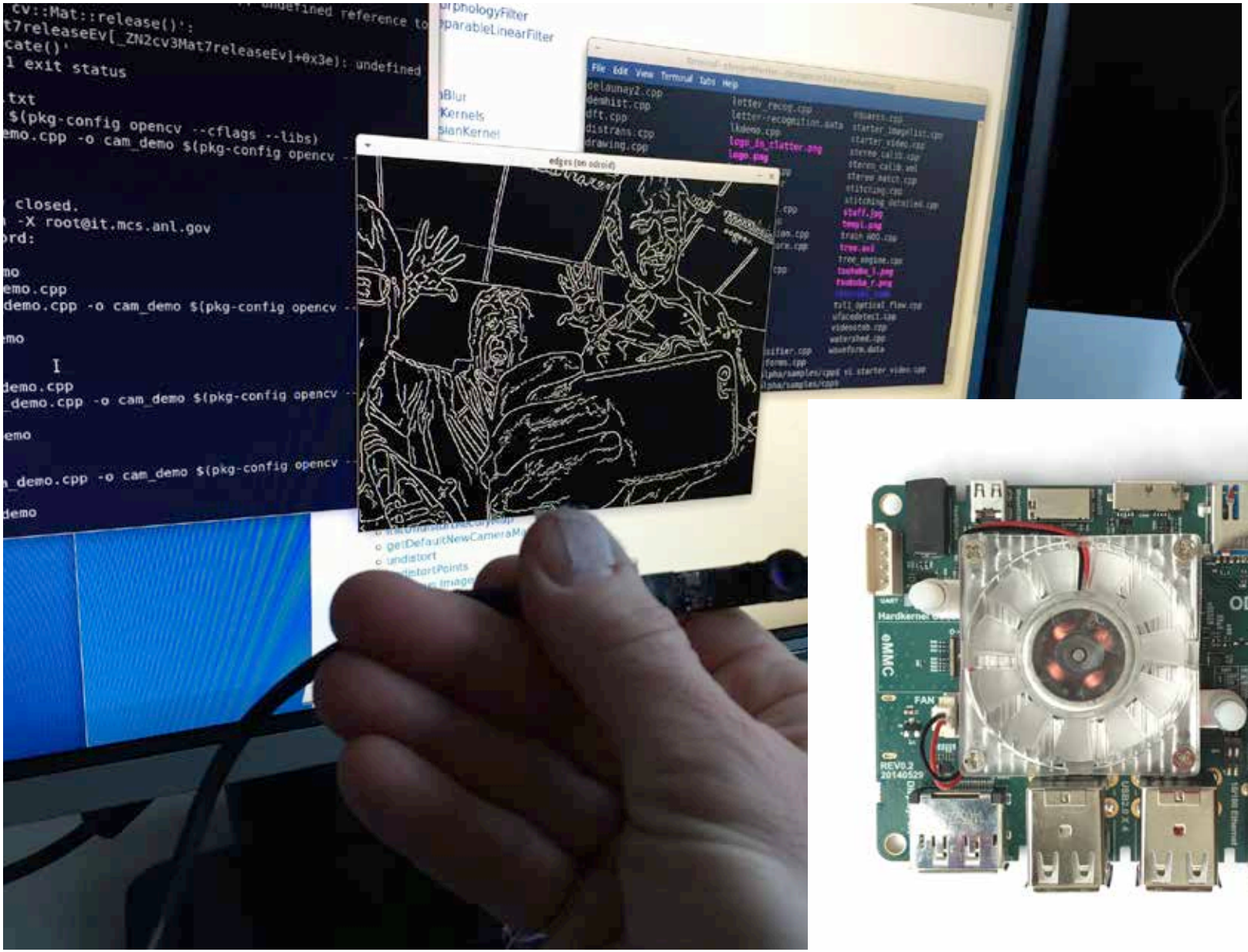
370 acres, owned/managed by Nature Conservancy and Northeastern IL Univ.
Registered as National Natural Landmark





al Laboratory / Northwestern University

In-Situ Analysis and Feature Recognition

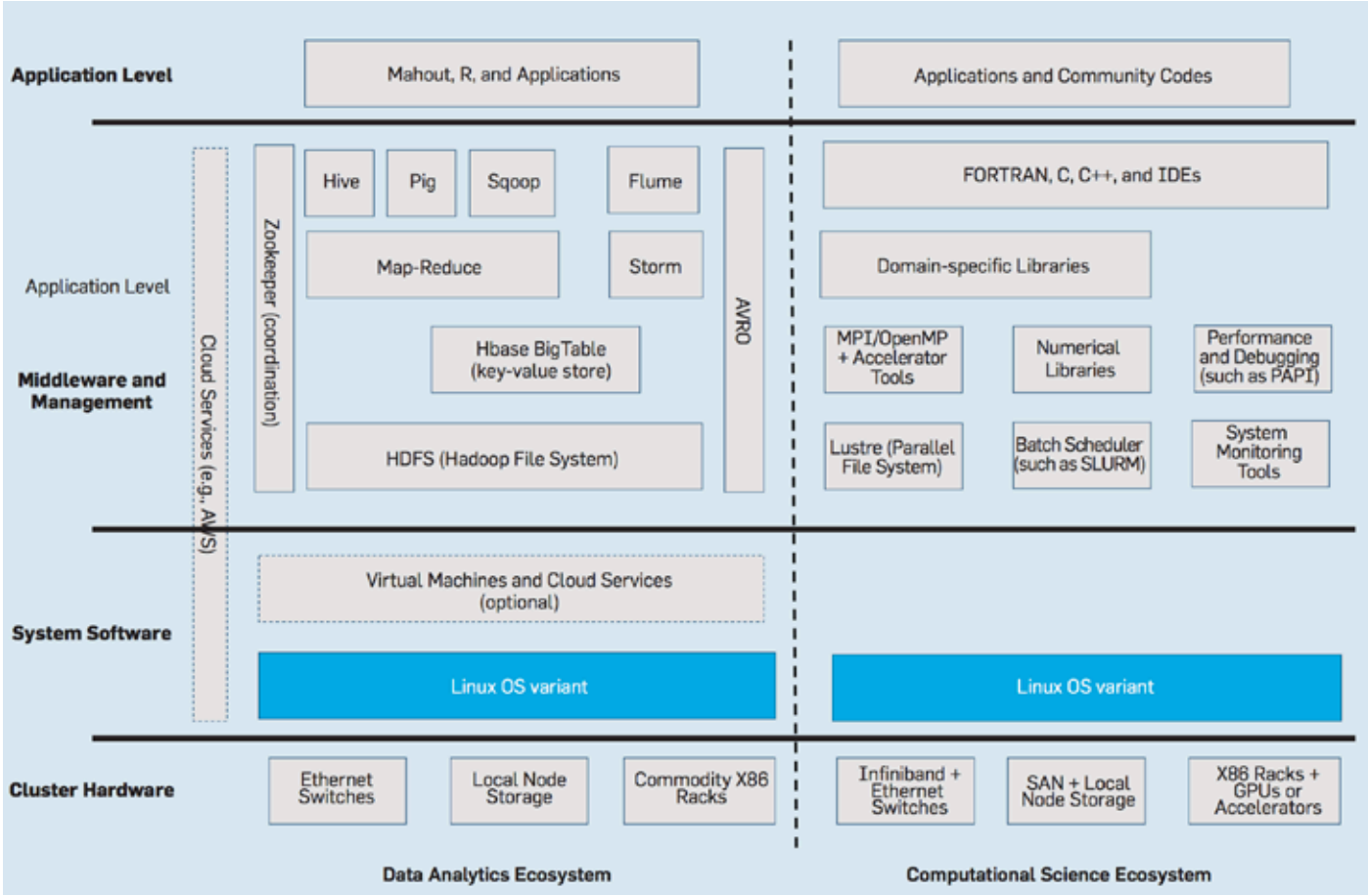



LinkNYC
NYC RESTAURANT WEEK
JAN 14 - FEB 8
NYCGO.COM

Google


sidewalk
LABS
Our mission is to improve life in cities for everyone through the application of technology to solve urban problems


BOTTLES CANS PAPER
RECYCLING ONLY
http://www.flickr.com/photos/firetransitcamera/
© 2012 J. Mc



Array of Things BDEC Needs:

- First-level in-situ analysis & cache
- Streaming Data to Cloud
- Event-driven HPC Simulations
- Multiple Data Source Integration
- Large complex data queries

Success Jail

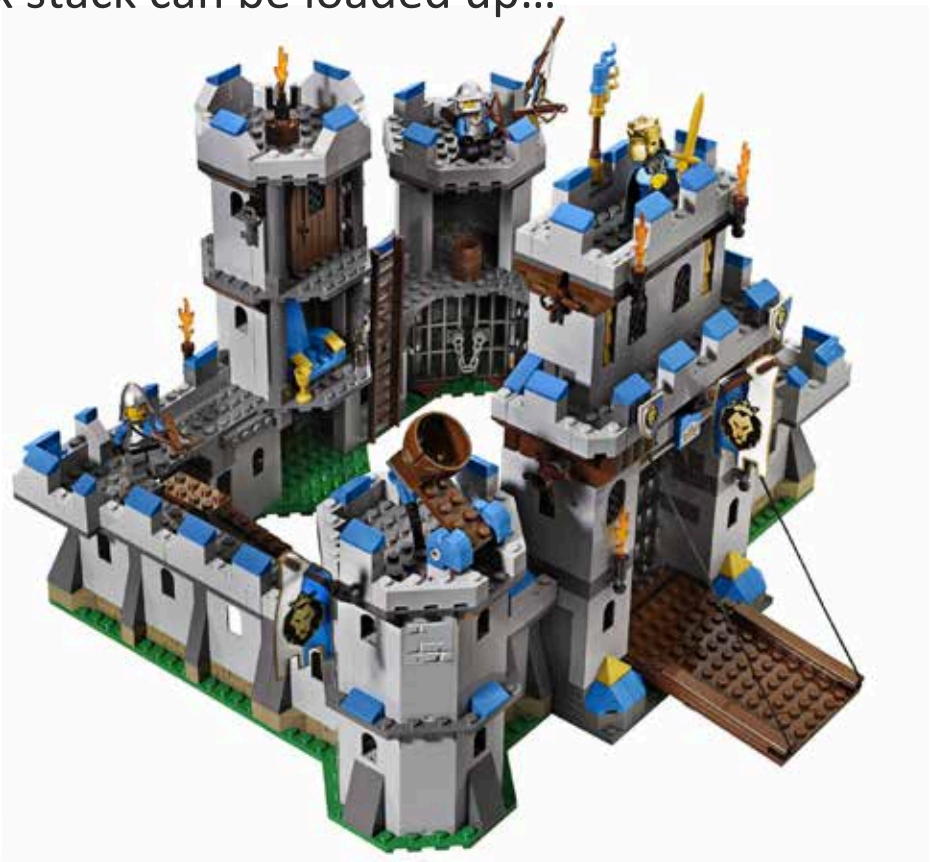
- HPC Systems have become very stable and predictable
- Dan Reed:
  - The questions remain the same, but the answers change
- How do you build a fast (economical) supercomputer?
  - Vectors
  - MPPs
  - Commodity CPUs
  - SOCs with embedded interconnect
- But what about ***operating?***

“Some people have to go to work and don't have time for all that”

- How do ***operate*** a fast supercomputer?
  - The answer has changed very little over 30 years
  - Submit a job, wait your turn, gain exclusive access
  - Focus on utilization for expensive resource (CPU)
  - Read data from parallel file-system, compute, write data back
  - Provide a single software stack for all users
  - Workflow? You can submit more jobs? That’s a workflow.
  - Reduce OS services to improve functionality of legacy block-synchronous computing

To Achieve Convergence: Revolt… Storm the Castle

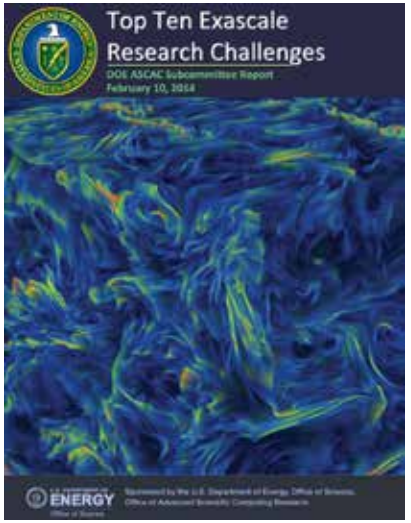
- Some new software technology and hardware helpful
- Operational practices and concepts must change
  - Then Geoffrey Fox stack can be loaded up…



Example: Why? For Science?

- **Trinity/NERSC-8:** ?

"The system shall provide correct and consistent runtimes. An application's runtime (i.e. wall clock time) shall not change by more than 3% from run-to-run in dedicated mode and 5% in production mode."



ASCAC Top 10 Research Challenges for Exascale

- "[…] power management [..] through dynamic adjustment of system balance to fit within a fixed power budget"
- […] Enabling […] dynamic optimizations […] (power, performance, and reliability) will be crucial to scientific productivity. "
- " […] Next-generation runtime systems are under development that support different mixes of several classes of dynamic adaptive functionality. "

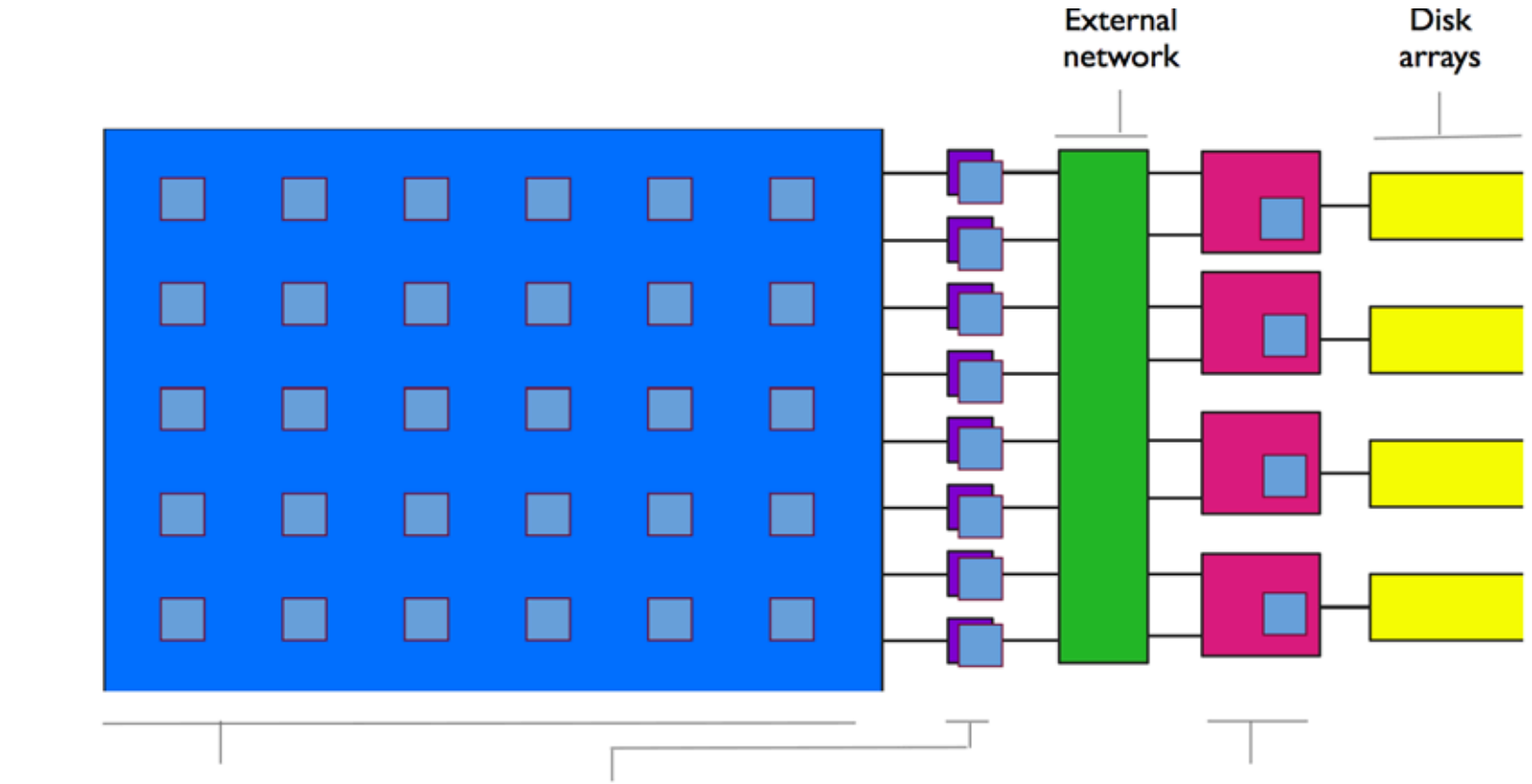
"dynamic" mentioned 43 times in 86 pg report

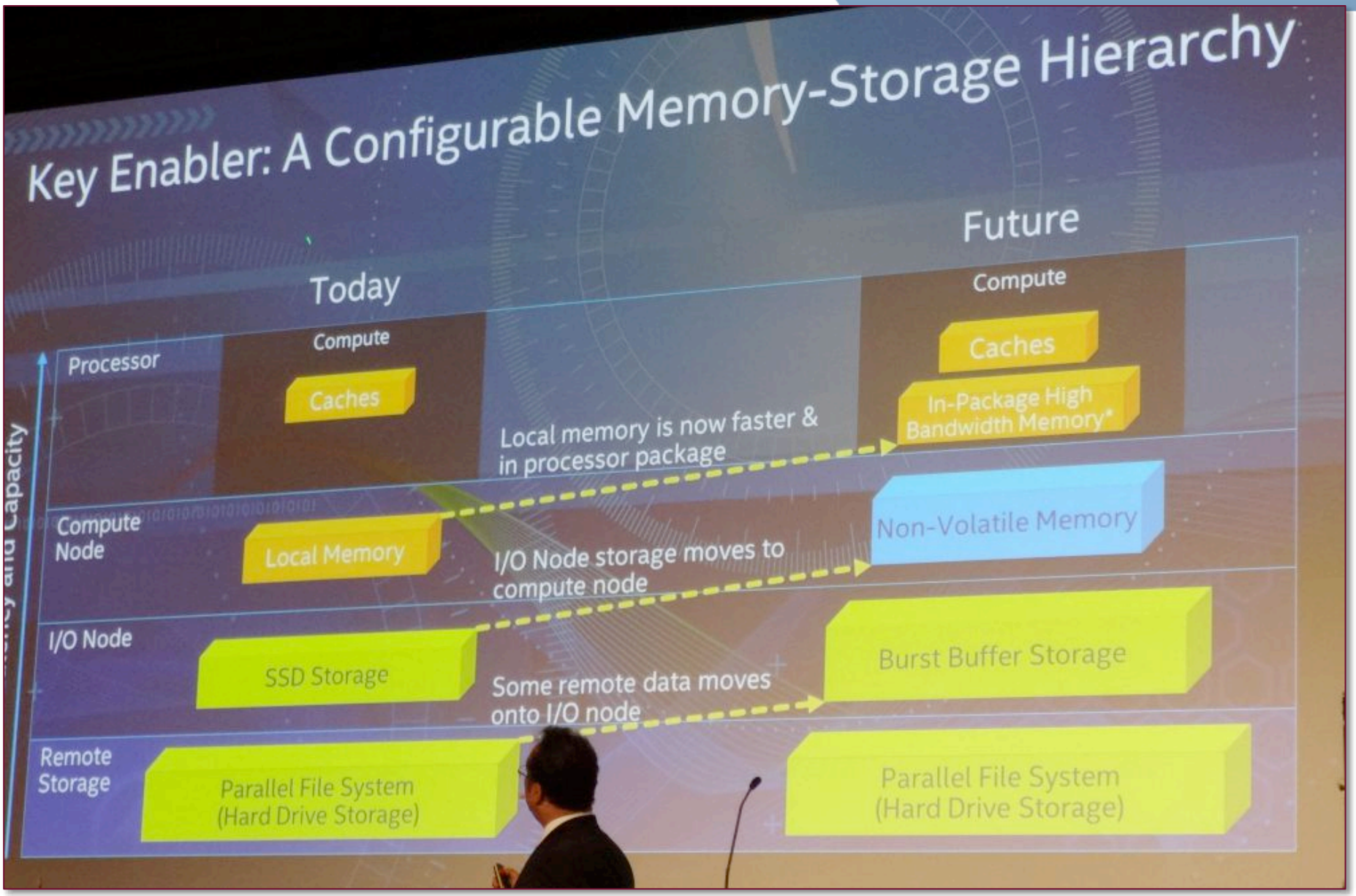
Convergence Principle #1
Allow Sharing

~~Submit a job, wait your turn, gain exclusive access~~
~~Focus on utilization for expensive resource (CPU)~~

Requires new system resource managers, provisioning, accounting, plumbing, etc.

Von Neumann Architecture is Recursive?





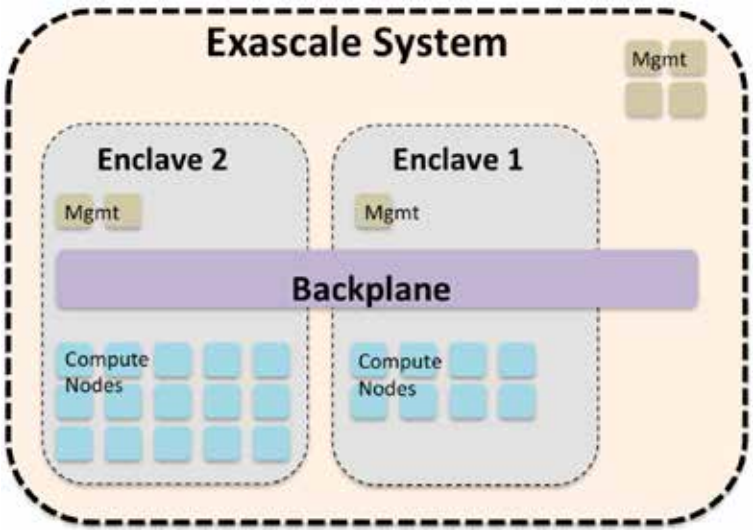
Raj Hazra, ISC, July 2015. Image from N. Hemsworth, "One Single System Architecture to Rule Them All," July 20, 2105.

Convergence Principle #2
Data and Compute Must Be Co-Located (Scheduled)

~~Read data from parallel file-system, compute, write data back~~

Requires new Operating System techniques, resource management, QoS guarantees

Exascale Operating System (With a Global View)



New abstractions & implementations

ANL: Pete Beckman, Marc Snir, Pavan Balaji, Rinku Gupta, Kamil Iskra, Franck Cappello, Rajeev Thakur, Kazutomo Yoshii
LLNL: Maya Gokhale, Edgar Leon, Barry Rountree, Martin Schulz, Brian Van Essen
PNNL: Sriram Krishnamoorthy, Roberto Gioiosa
UC: Henry Hoffmann
UIUC: Laxmikant Kale, Eric Bohm, Ramprasad Venkataraman
UO: Allen Malony, Sameer Shende, Kevin Huck
UTK: Jack Dongarra, George Bosilca, Thomas Herault

Convergence Principle #3
Support Containers and VMs

~~Provide a single software stack for all users~~

~~Reduce OS services to improve functionality of legacy block-synchronous computing~~

Requires new Operating System technologies to support containerization, QoS, and security

Convergence Principle #4
Get a Real Workflow System

~~Workflow? You can submit more jobs? That's a workflow.~~

Requires a very sophisticated workflow system that understands dependencies, cost models, optimization

Summary: How To Storm The Castle… (and bring about The Convergence)

- Allow Sharing
  - Increase interactivity, allow extremely long-running jobs
- Data and Compute Must Be Co-Located (Scheduled)
  - We are getting NVRAM everywhere. We must build new resource managers that can handle persistent data left behind and favored job positioning
- Support Containers and VMs
  - Not that hard, but we have to add some hardware support for security….
- Get a Real Workflow System
  - Hmmmm, if only we knew people who were exploring this area….

Finally: Load up ***Cool Big Data software stacks….***

Questions?