

The Role of Dedicated Data Computing Centers in the Age of Cloud Computing

CHEP 2016 – San Francisco

Tony Wong

Brookhaven National Laboratory

BNL



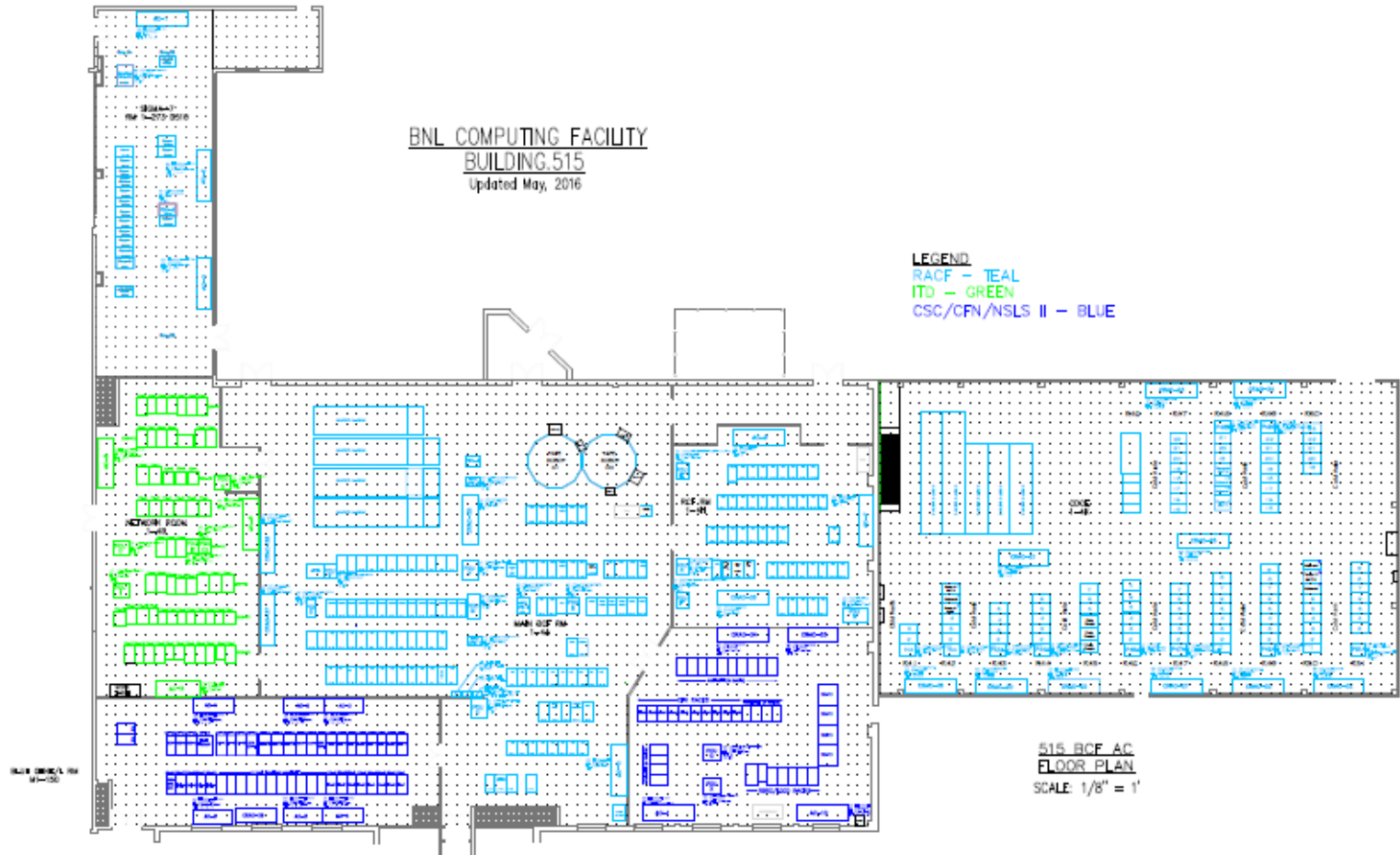
Computing @ BNL (1)

- RHIC-ATLAS Computing Facility (RACF) supports RHIC experiments and US ATLAS
- Computing center of BNL's Computation Science Initiative (CSI) is called Scientific Data & Computing Center (SDCC)
- SDCC is housed and operated by RACF staff
- SDCC focus is on support for HPC activities

Computing @ BNL (2)

- Existing data center space mostly devoted to RHIC and ATLAS is nearly full
 - 15,000 ft² (~1,400 m²), including new expansion space built in 2009
 - ~2.3 MW of UPS power
- HPC-centric existing space (~2,500 ft² and ~500 kW of UPS power) not sufficient
- Little space and power left for expansion to support new programs at BNL
 - Center for Functional Nanomaterials (CFN)
 - Computational Science
 - Others

Data Center



Existing Data Center Is Full



Computing Resources

- Dedicated HTC resources
 - Tier 0 for RHIC computing
 - U.S. Tier 1 and Tier 3 for ATLAS
 - Other (LBNE, LSST, etc)

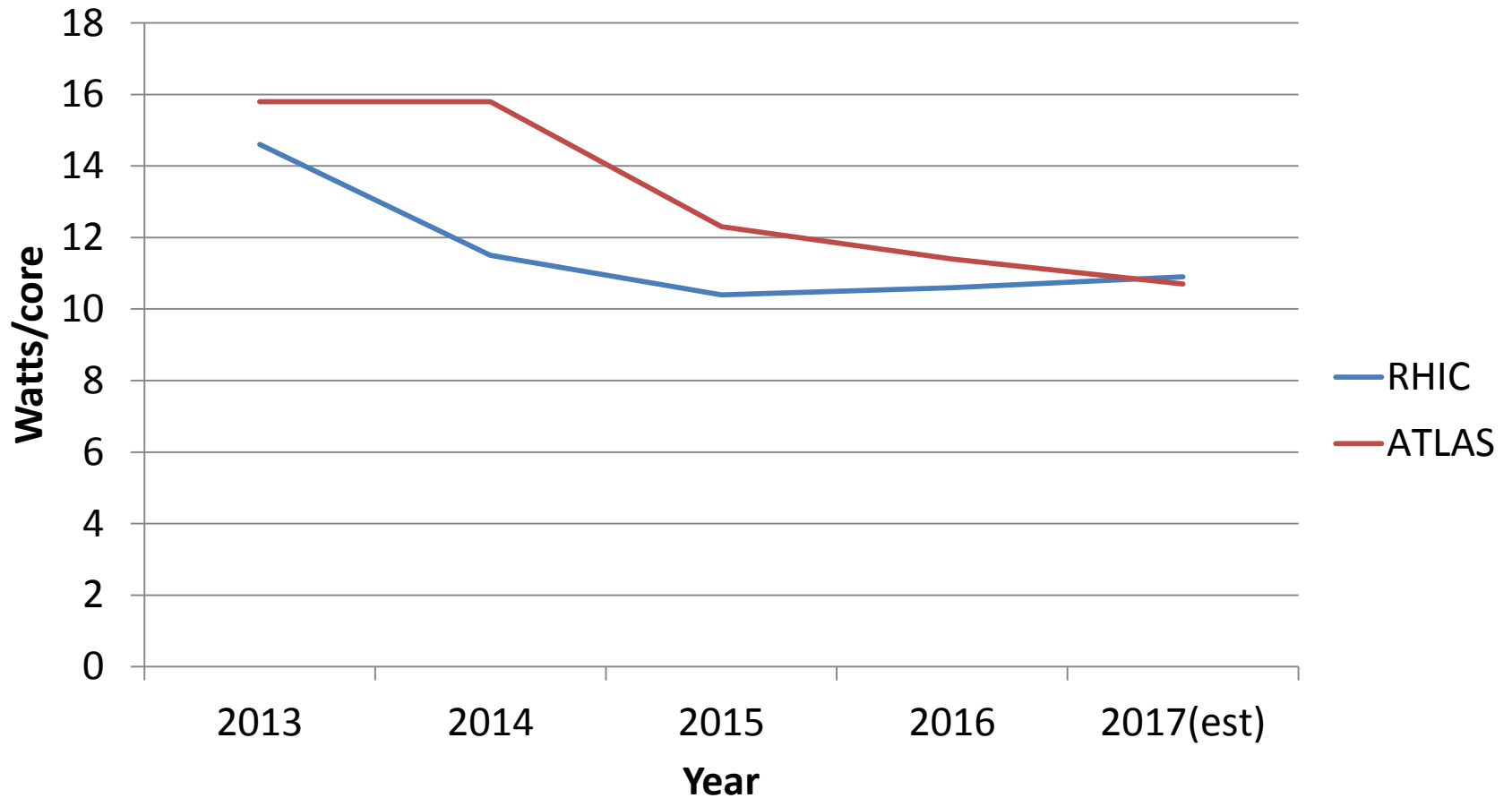
Year	Servers	Cores	HS06	Distributed Storage on worker nodes (in PB)
2013	2.2k	23.7k	380k	16.0
2016	2.2k	61.0k	550k	26.3

- Significant HPC resources deployed (~15k cores)

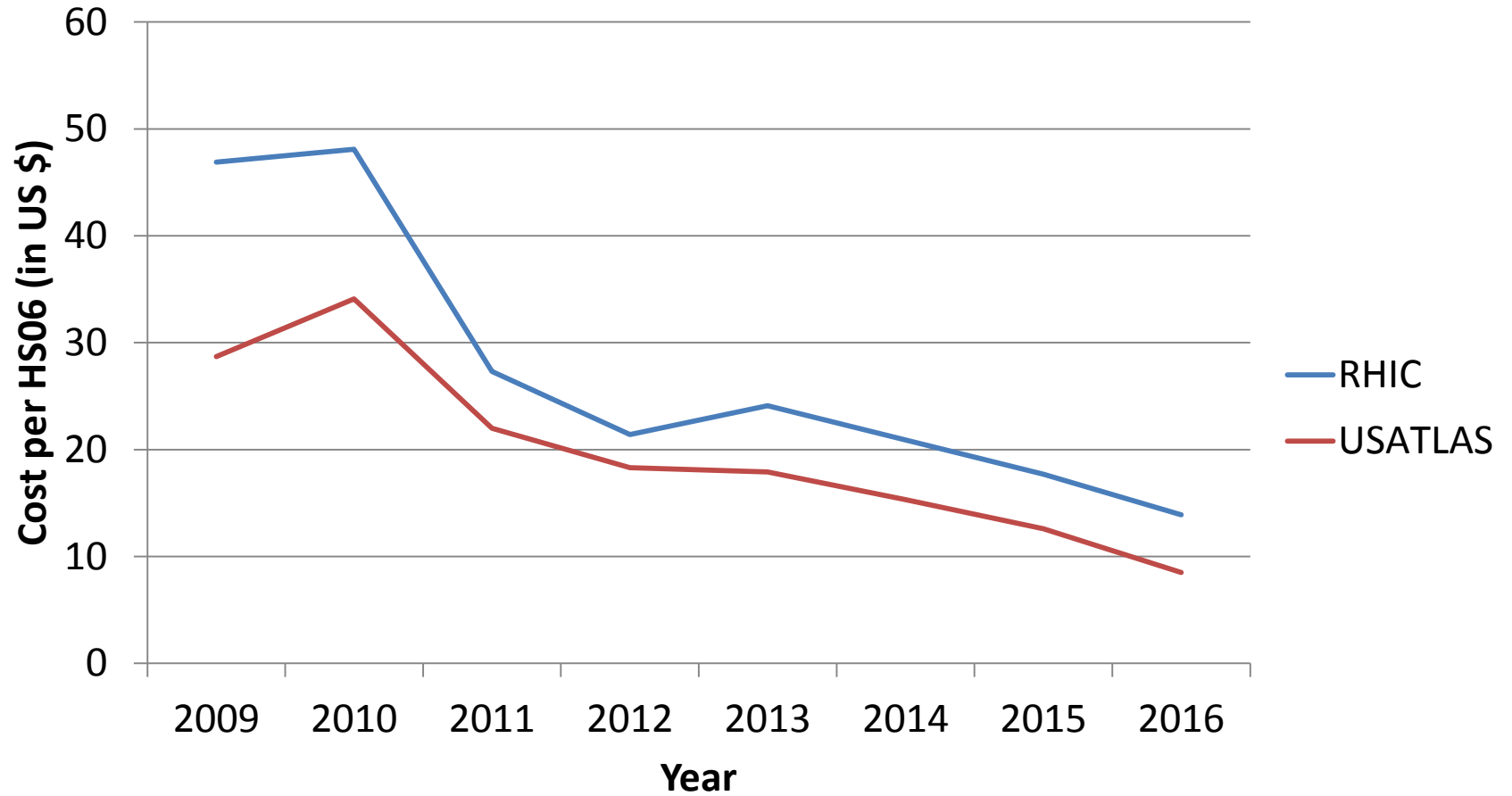
Cluster	Servers	Cores	Power (in KW)
IC	108	3888	120
KNL	144	9216	72
Other (legacy)	176	1856	27

Good News

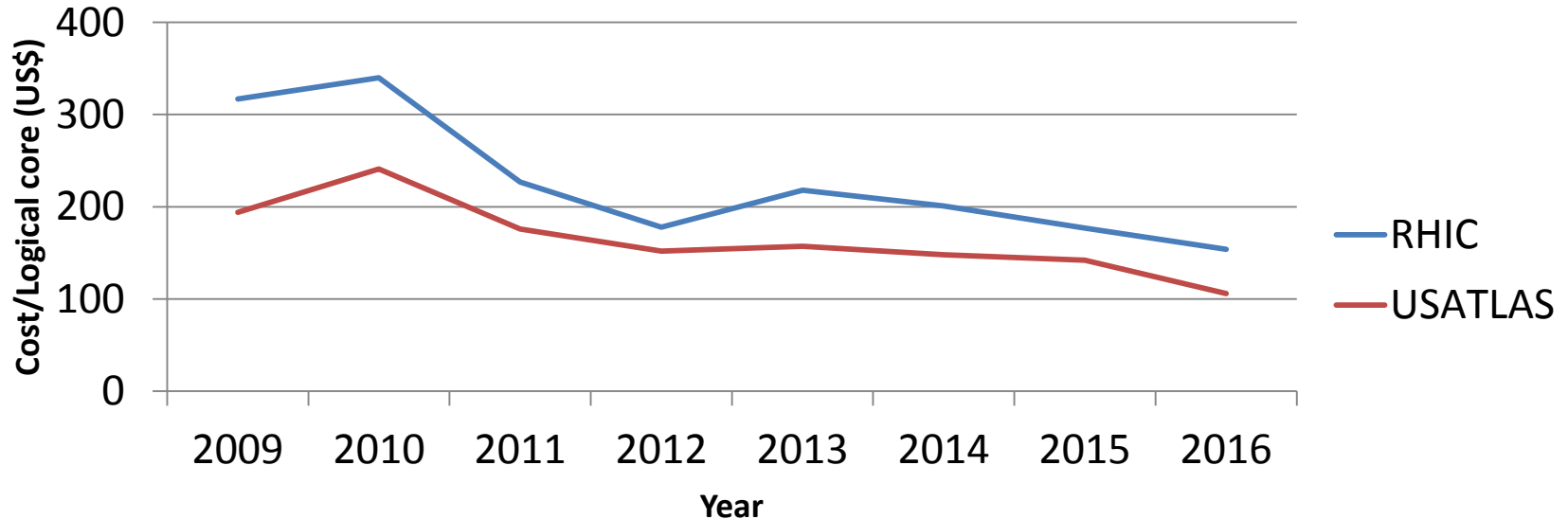
Cpu's are more power-efficient



Cost per HS06 is trending down



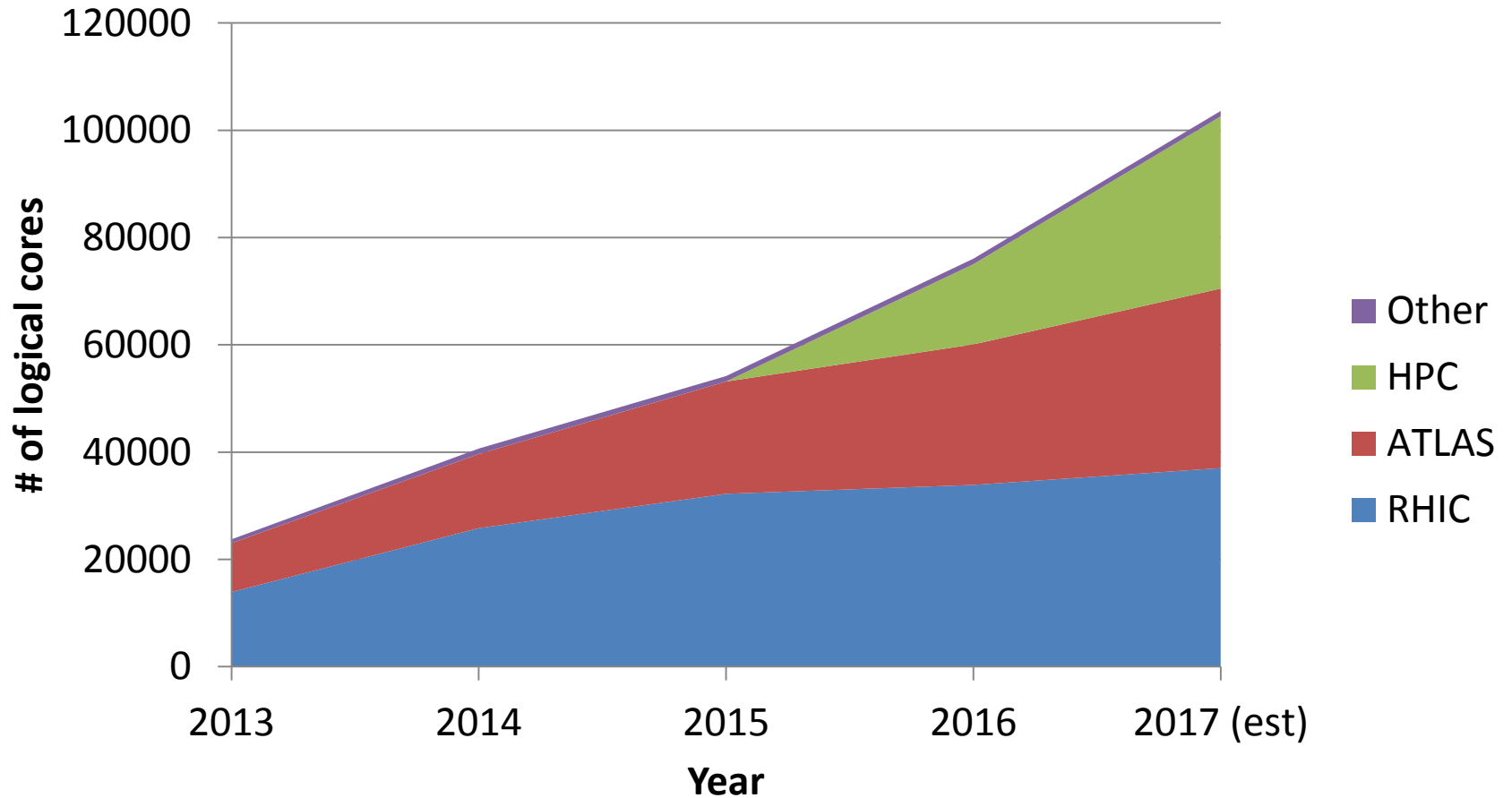
Hardware cost is trending down



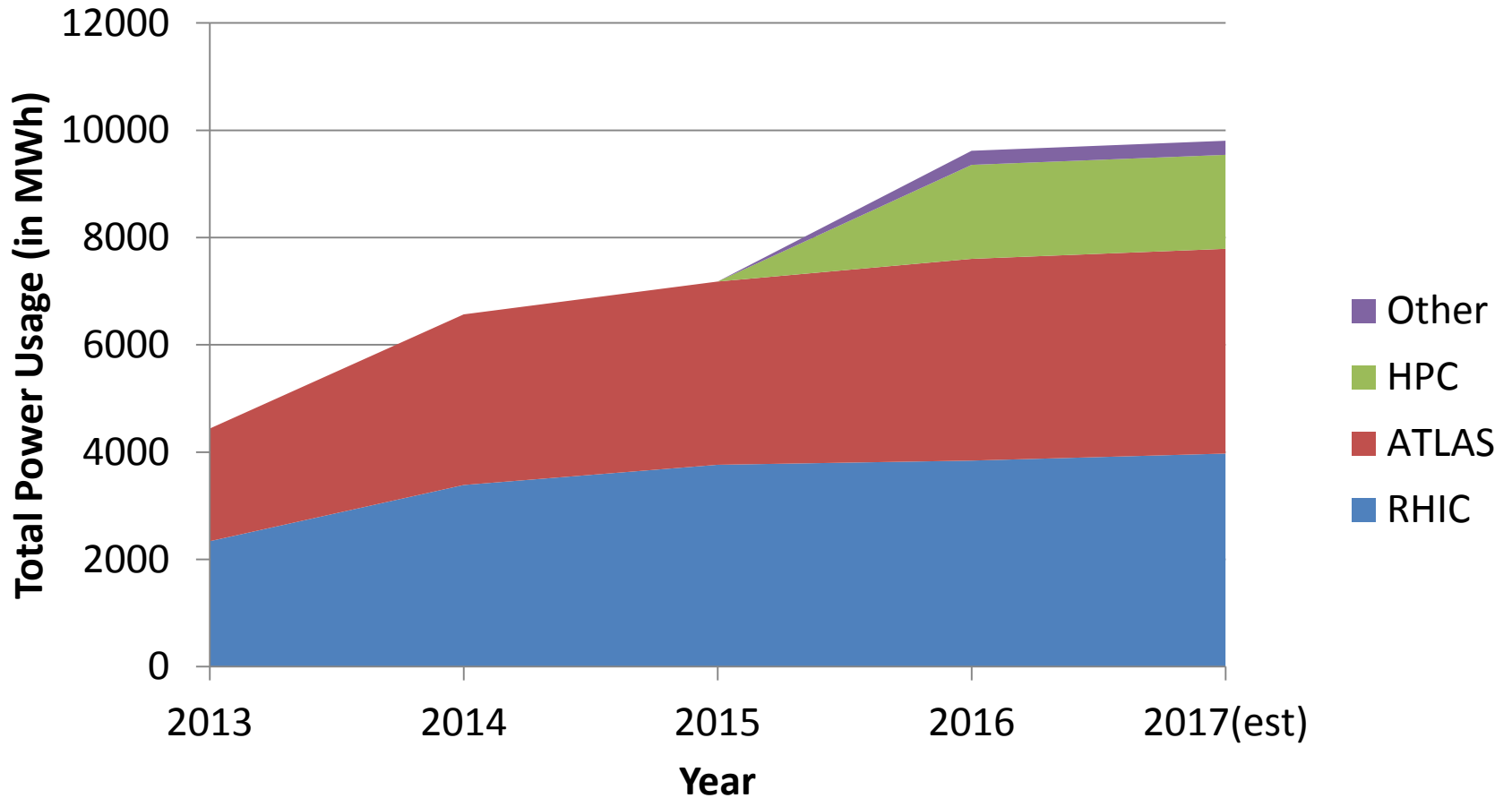
- Standard 1-U or 2-U servers
- Includes server, rack, rack pdu's, rack switches, all hardware installation (does not include network cost)
- Hardware configuration changes (ie, more RAM, storage, etc) not decoupled from server costs → partly responsible for fluctuations

Bad News

Fast-rising computing requirements



Growing power usage



Proposed New Data Center

- Review process underway
 - CD-0 granted Fall 2015 (science case)
 - CD-1 review Summer 2016 (alternative analysis)
- New Data Center construction timeline (if approved)
 - Preliminary design in 2017
 - Construction begins in 2018
 - Most realistic scenario indicates occupancy in late 2021
 - Contingency plans for temporary space 2017-2021 to accommodate HPC growth and any other HTC-centric programs (ie, LSST, DUNE, etc)

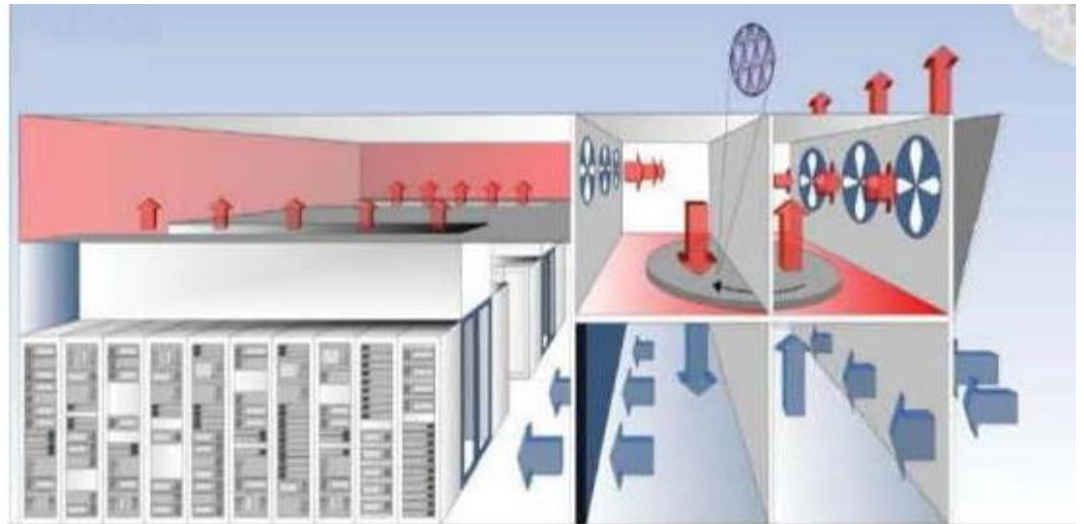
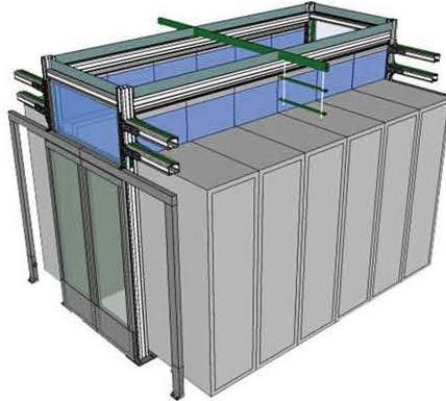
New Data Center Profile

- 25,000 ft² (~2,320 m²) of usable space
- 2.4 MW of UPS power on day 1 (expand up to 6 MW in future)
- PUE of 1.2 to 1.4 (mandated by DOE)
- Shared facility for ATLAS, RHIC, CSI, Photon Science
- Natural air-cooled supplemented by redundant chillers
- Hot-aisle containment

CFR – Conceptual Design

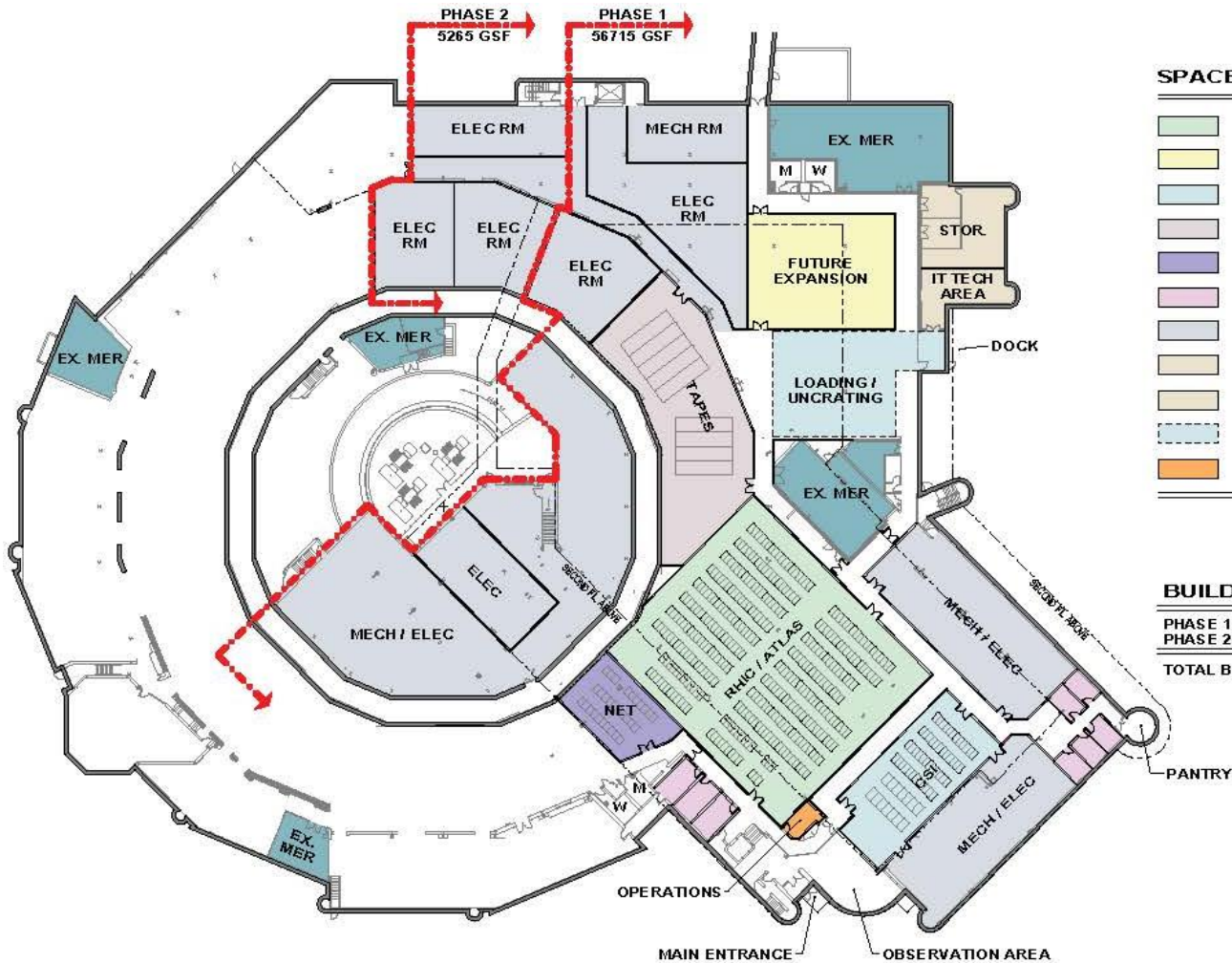
Data Center Systems

- Aisle Containment
 - Integrated provisions for air path containment, cable trays, and power raceways.
 - Required due to anticipated power density.



Theory behind rotary heat exchanger technology

CFR – Schematic Floor Plan



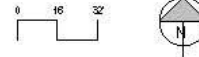
SPACE LEGEND

	PROGRAM NET AREA (SGF)	DELIVERED AREA (SGF)
RHIC / ATLAS	1080	3271
FUTURE EXPANSION	350	2688
CSI	450	2238
TAPES	300	4159
NETWORK EQUIPMENT	900	1195
OFFICES	830	1195
MECH/ELEC	2020	2894
IT TECH AREA	500	643
STORAGE	200	200
LOADING / UNCRATING	500	279
OPERATIONS	191	132
TOTAL NET AREAS	4390	5495

BUILD-OUT AREAS

	AREA (GSF)
PHASE 1	56715
PHASE 2	5265
TOTAL BUILD-OUT AREA	61980

DATE:	01 AUGUST 2016
PROJECT:	CORE FACILITY REVITALIZATION (CFR) BUILDING 725
DISCIPLINE:	ARCHITECTURE
SHEET:	2.4.1.2 FIRST FLOOR PLAN
 	



External Considerations

- DOE mandate to prioritize “Cloud” as alternative to building new data centers
- Current budgetary realities and program requirements have compelled the HENP community to evaluate off-site alternatives, independent of DOE mandate
- Commercial providers (Amazon, Google) offer increasingly price-competitive cloud services
- Virtual (non-profit) organizations (ie, OSG) are harnessing the compute power of non-dedicated (HTC and HPC) resources

Alternative Analysis for CD-1

- Four scenarios considered
 1. Do nothing
 2. Utilize existing BNL facilities
 - a. Renovate current data center
 - b. Re-purpose another building
 3. Build new facility
 4. Use cloud resources
- Option 1 not viable given the growth of computing resources and option 2a not possible with concurrent operations (unacceptably long downtimes)
- Total cost (building + 25-yr operational lifetime) for option 3 is ~20% higher than option 2b

Alternative Analysis for CD-1

- Four scenarios considered
 1. Do nothing
 2. Utilize existing BNL facilities
 - a. Renovate current data center
 - b. Re-purpose another building
 3. Build new facility
 4. Use cloud resources
- Compare two most cost-effective solutions (options 2b and 4) on a hypothetical 3-yr deployment and operations scenario
- Several assumptions made to simplify calculations
 - Local hosting (power, cooling, staff, etc) costs remain constant
 - Future requirements do not deviate from forecast estimates
 - Tape storage (capital and operations) not included—even though it is essential component of archival storage at RACF

Cloud Resources (1)

- Estimate based on AWS prices as of July 2016
 - Based on projected disk storage needs over next 3 years
 - Increment capacity by 2.33 PB/yr to meet storage requirements
 - Table below only shows cost of standard storage (frequent access)
 - No cost for data transfer to Amazon S3
 - Cost of data transfer out from S3 to Internet is 2-3x cost of storage (\$0.05 to \$0.09/GB) not included in the table below

Year	Cost per PB/month (in US\$)	PB/yr	Cost/yr (in US\$)	Cumulative Cost (in US\$)
1	27.5k	2.33	770k	770k
2	27.5k	2.33	1,540k	2,310k
3	27.5k	2.33	2,310k	4,620k

Cloud Resources (2)

- Estimate based on EC2 spot prices as of July 2016
 - Assume 10% of computing needs over next 3 years use cloud resources
 - Use EC2 equivalent (20% improvement/yr) to RACF 5,000 cores
 - EC2 equivalent also assumes 80% cpu efficiency at the RACF
 - Based on c4.large instance
 - Assume 100% job efficiency with spot pricing in table below for simplicity (note: BNL experience indicates otherwise)

Year	Spot c4.large (\$/yr)	Equivalent cores	Cost/yr (in US\$)	Cumulative Cost (in US\$)
1	151.5	3,077	466k	466k
2	151.5	2,564	389k	855k
3	151.5	2,137	324k	1,179k

In-House

- Disk storage
 - Total cost of ownership (TCO) for 7 PB is **~\$1.3M** (including hardware and data center operations)
- Computing
 - TCO depends on lifetime of resource
 - Cost is amortized over lifetime

Model	\$/core (in US\$)	Data Center charges/yr (in US\$)	Total cost (in US\$)
3-yr	70.7	78k	588k
5-yr	63.6	78k	708k

Cloud vs. In-House

- In-House is more cost-effective for both computing and disk storage in a 3-yr scenario
 - \$3.3M less for disk storage
 - \$0.6M less for computing
- Staff costs not significantly different with either solution
- Cost of data movement not estimated for either solution
- Cost of data replication not included for in-house or cloud solution
- So where does cloud computing fit at BNL?

Cloud Activities at RACF

- BNL cloud cluster
 - Re-purposed RACF hardware with OpenStack provisioning
 - Targets user base with modest computing requirements
 - Estimated to cost \leq \$0.02/node-hr
- OSG & other remote access
 - Integrated into RHIC and ATLAS Tier 1 resource allocation for several years
- Amazon & Google
 - Continued interest & evaluation

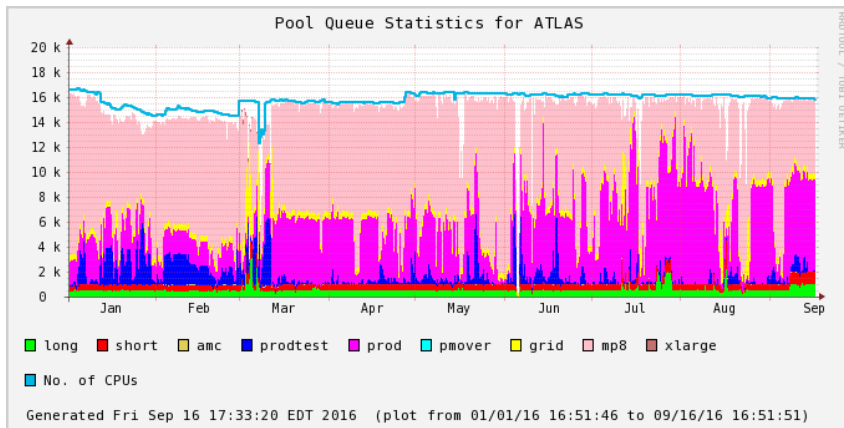
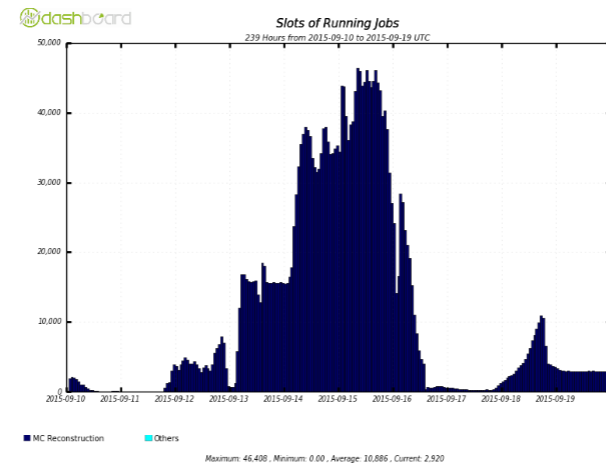


Figure 3. Cores, Sept 10 - Sept 20, 2015 (jobs=cores/8)



Computing Profiles

- HTC (generally speaking)

- Simulation

- minimal or no local dependencies
 - long-running, cpu-bound jobs



Budget-friendly
cloud workload

- Analysis and Data Processing

- distributed systems with relatively 'slow' networks
 - loosely coupled, system-bound jobs

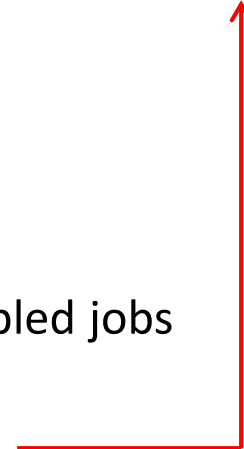
- HPC (generally speaking)

- Low-latency interconnect

- high-performance parallel file system
 - batch system optimized for multi-node, tightly-coupled jobs

- Multi-core serial processing

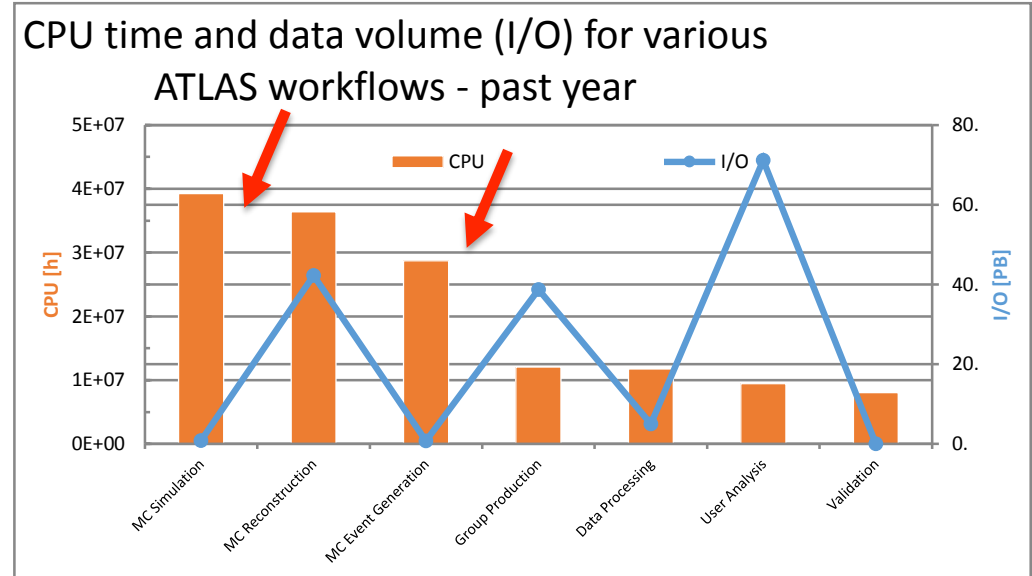
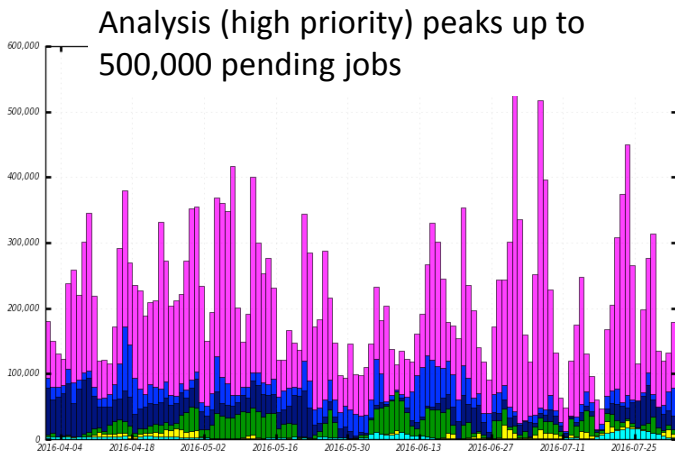
- minimal or no local dependencies
 - short-duration, high-frequency, cpu-bound jobs



Cloud Services an Alternative?

(slide provided by Eric Lançon)

- Not all ATLAS workflows are suitable for Cloud Services
 - Only high CPU, low I/O and low priority **workflows**
 - Others workflows:
 - Too much I/O
 - Too high priority to benefit from Spot market prices



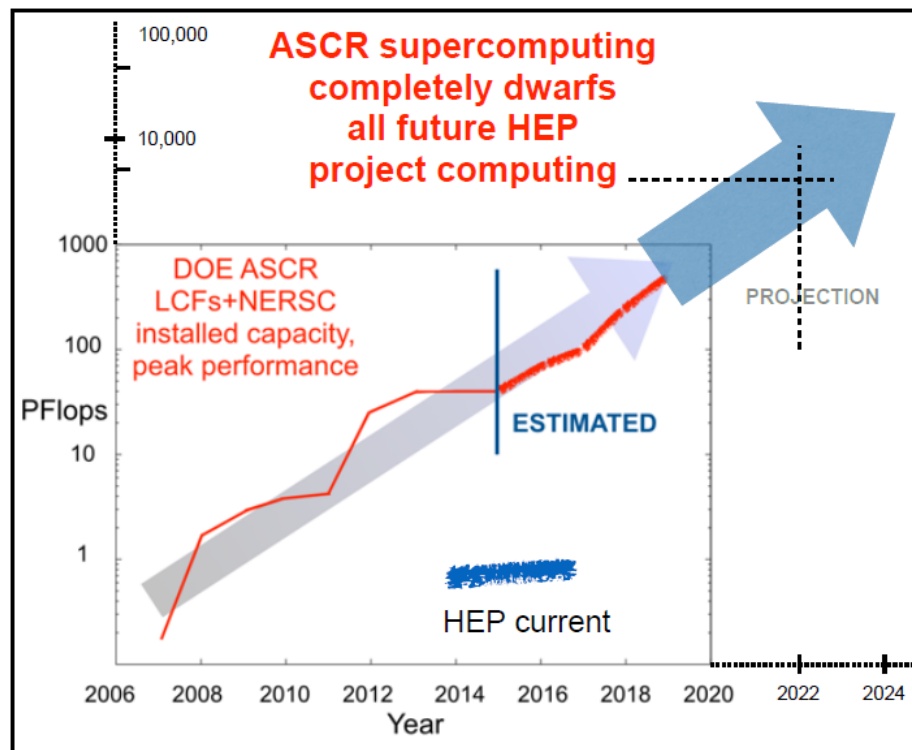
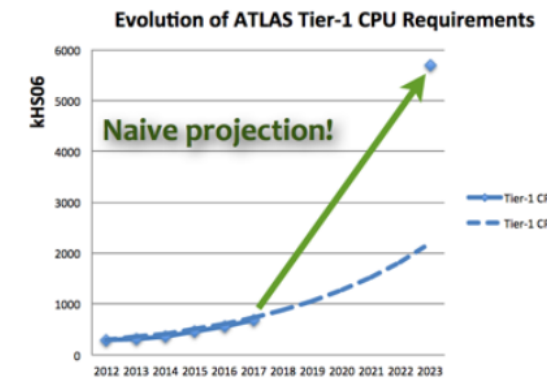
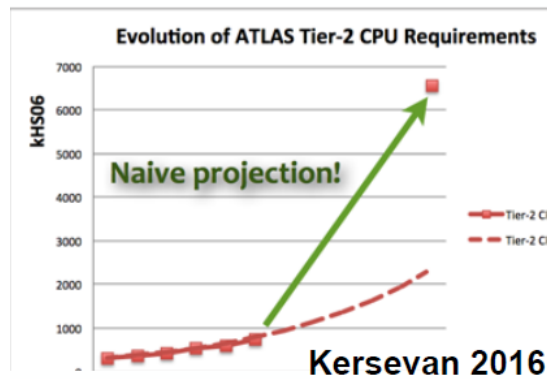
- RACF is designed to meet these requirements

Future HEP Requirements

(from Salman Habib's (ANL) presentation at NYSDS 2016)

HEP Computing Requirements for Energy Frontier

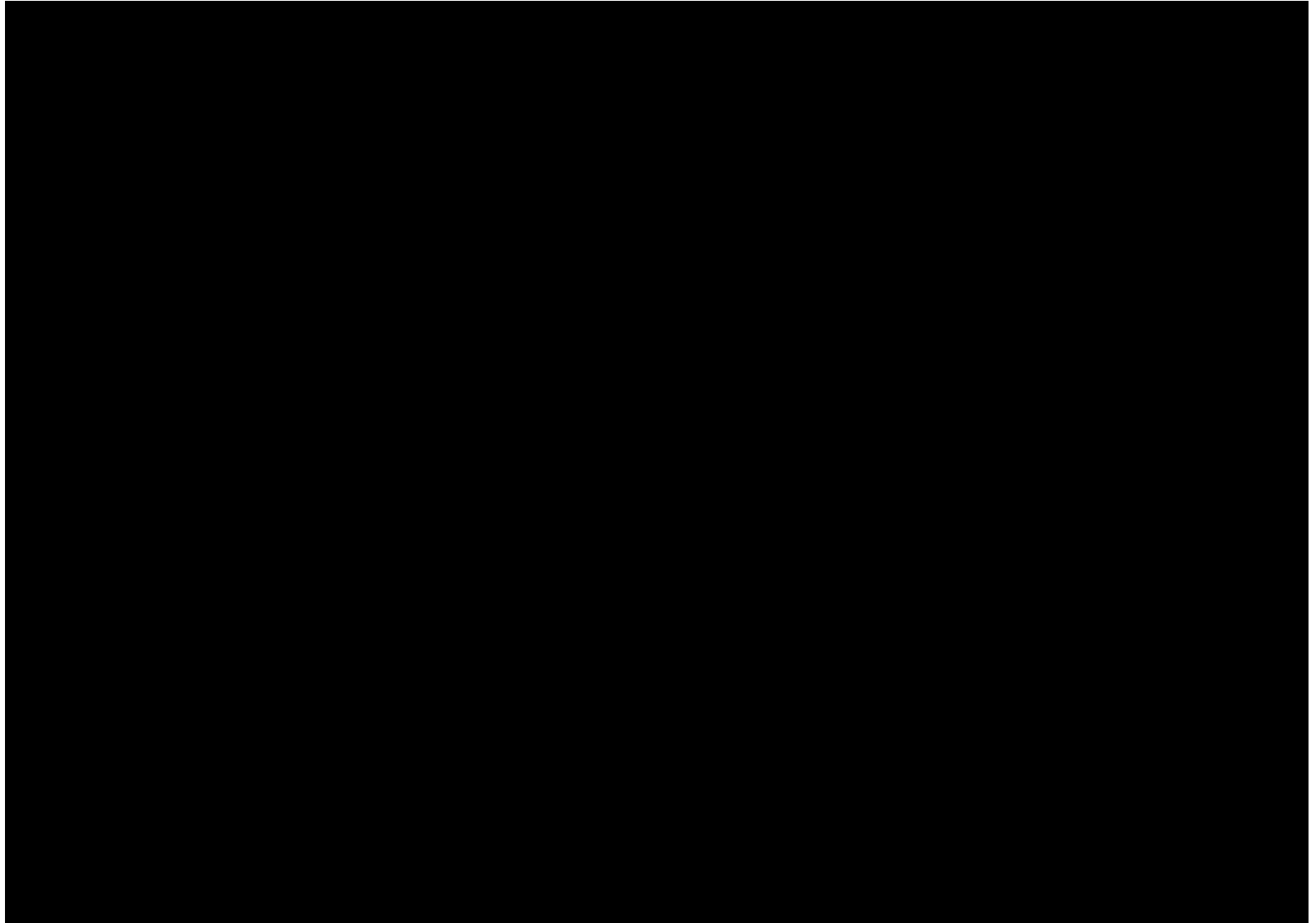
- HEP Requirements in computing/storage will scale up by ~50X over 5-10 years
 - Flat funding scenario fails — must look for alternatives!



Cloud vs. in-House cost evolution

- BNL presentation at CHEP 2013 in Amsterdam(<http://iopscience.iop.org/1742-6596/513/6/062053>)
 - Computing
 - \$0.013/hr (m1.medium spot instance)
 - **\$0.02/hr (RACF)**
 - \$0.12/hr (m1.medium on-demand instance)
 - Storage
 - \$0.05/GB/month
- Current AWS costs (as of July 2016)
 - Computing
 - \$0.017/hr (c4.large spot instance)
 - **\$0.015/hr (RACF)**
 - \$0.105/hr (c4.large on-demand instance)
 - Storage
 - \$0.0275/GB/month
- Note: switched to c4.large instance to match current requirements

New HPC Cluster(s)



Summary

- In-house cost-competitive with cloud resources
 - True over past ~4 years – confident it will hold true over 25-yr lifetime of data center
 - Irreducible cost of hardware makes up ~70% of Total Cost of Ownership – hard floor to any further competitive gains at BNL or elsewhere
- Access to cloud resources still important
 - Upcoming HEP computing/storage requirements cannot be met without “external” contributions
 - In-house competitiveness depends on volatile factors (cost of electrical power, infrastructure support, etc) and cannot be taken for granted as enduring advantages
 - Motivates the development of mechanisms and models for cost-effective access, such as event server to use AWS spot pricing