Operating Dedicated Data Centers – Is It Cost-Effective?

CHEP 2013 - Amsterdam

Tony Wong - Brookhaven National Lab

Motivation

- Traditional HENP computing models mostly relied on dedicated data centers until recently
- Budgetary realities (and economies of scale) compel the HENP community to evaluate alternatives
- For-profit providers (Amazon, Google) offer cloud services
- Virtual organizations (OSG, EGI, etc) are harnessing the power of non-dedicated resources

RACF Context at BNL

- Dedicated resources
  - Tier0 for RHIC computing
  - U.S. Tier 1 for ATLAS
  - Two (BNL, Wisconsin) Tier 3 for ATLAS
  - Other (LBNE, LSST, etc)
- Over 2,200 servers, 23,000 physical cores and 16 PB of worker node-based storage
- Robotic storage and other disk-based distributed storage services

Usage Profiles

- Monte Carlo
  - minimal local dependencies
  - long-running, cpu-bound, low I/O requirements
- Analysis
  - some local dependencies
  - Variable length and I/O requirements
- Interactive
  - significant local dependencies
  - short-running, high I/O requirements

Amazon EC2

	Type	ECU	RAM (GB)	Storage (GB)	Network I/O	Cost/hr (US$)
spot	m1.small	1	1.7	160	low	0.007
spot	m1.medium	2	3.75	410	moderate	0.013
On-demand	m1.medium	2	3.75	410	moderate	0.12

- Full details at aws.amazon.com/ec2/pricing.
- Linux virtual instance
  - 1 ECU = 1.2 GHz Xeon processor from 2007 (HS06 ~ 8/core)
  - 2.2 GHz Xeon (Sandybridge) in 2013 → HS06 ~ 38/core
- Pricing is dynamic and region-based. Above prices were current on August 23, 2013 for Eastern US.

BNL Experience with EC2

- Ran ~5000 EC2 jobs for ~3 weeks (January 2013)
  - Tried m1.small with spot instance
  - Spent US $13k
- Strategy
  - Declare maximum acceptable price, but pay current, variable spot price. When spot price exceeds maximum acceptable price, instance (and job) is terminated without warning
  - Maximum acceptable price = 3 x baseline → $0.021/hr
- Low efficiency for long jobs due to eviction policy
- EC2 jobs took ~50% longer (on average) to run when compared to dedicated facility

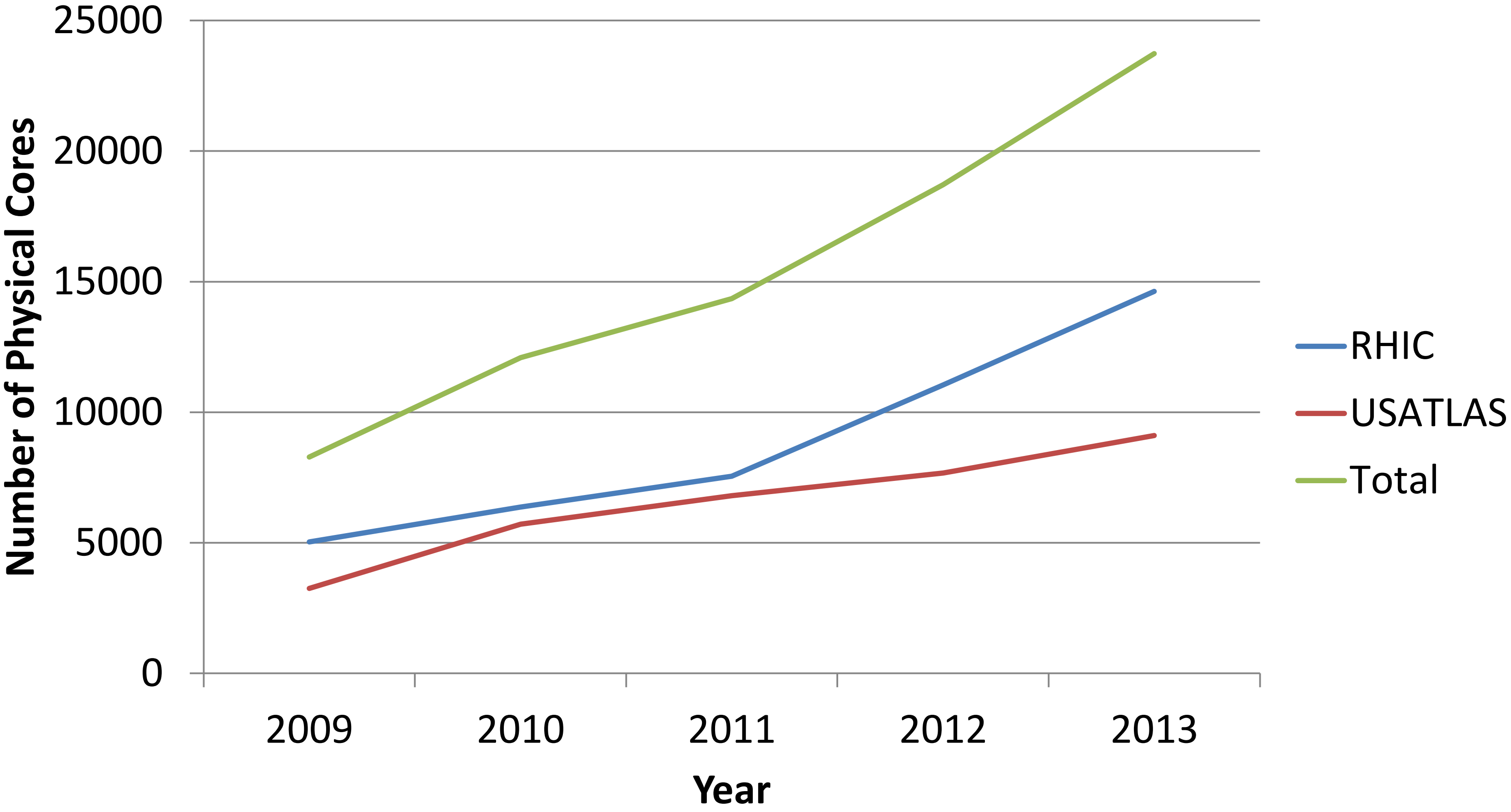
Google Compute Engine (GCE)

- Standard Instance type (equivalent to EC2's On-Demand)
  - Linux (Ubuntu or CentOS)
  - $0.132/hr (1 virtual core, 3.75 GB RAM, 420 GB storage) → similar to EC2's on-demand m1.medium
  - Evaluation
    - 458k jobs (mix of evgen, fast sim and full sim)
    - Ran for 7 weeks (March-April 2013)
    - Custom image based on Google's CentOS 6
    - Sergey Panitkin's presentation—ATLAS Cloud Computing R&D.

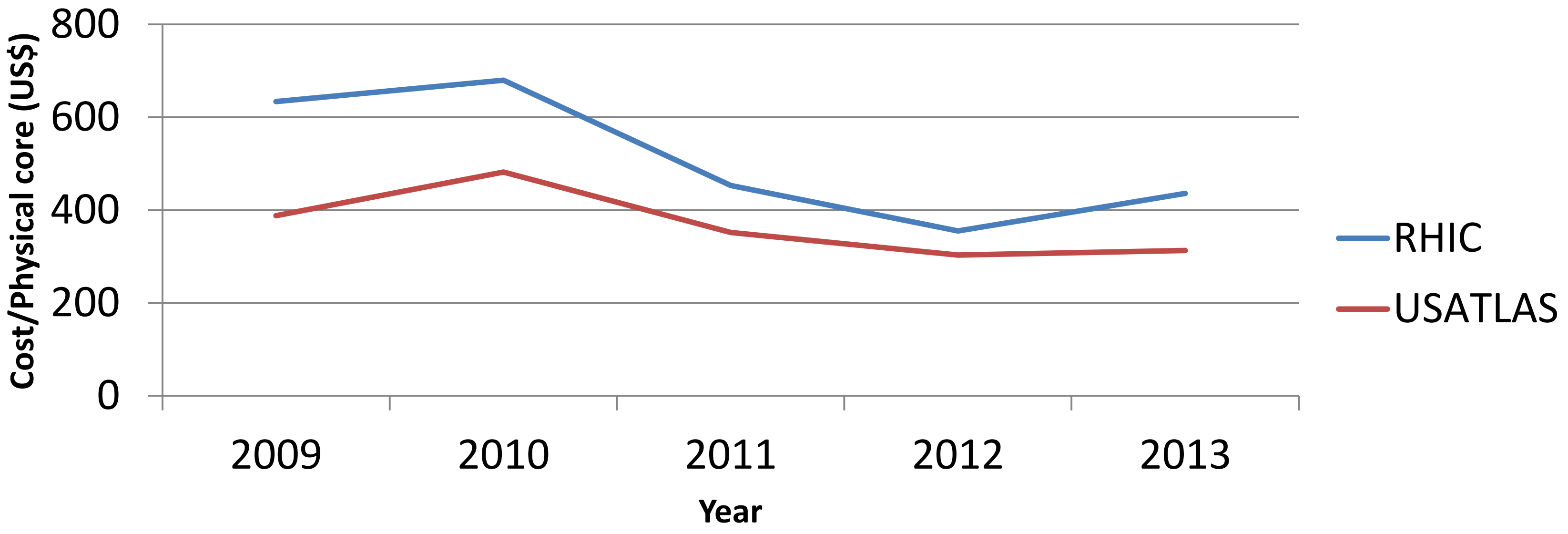
Costs of Dedicated Facilities

- Direct Costs
  - Hardware (servers, network, etc)
  - Software (licenses, support, etc)
- Indirect Costs
  - Staff
  - Infrastructure
    - Power
    - Space (includes cooling)
  - Other?

Growth of RACF Computing Cluster



Server Costs



- 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

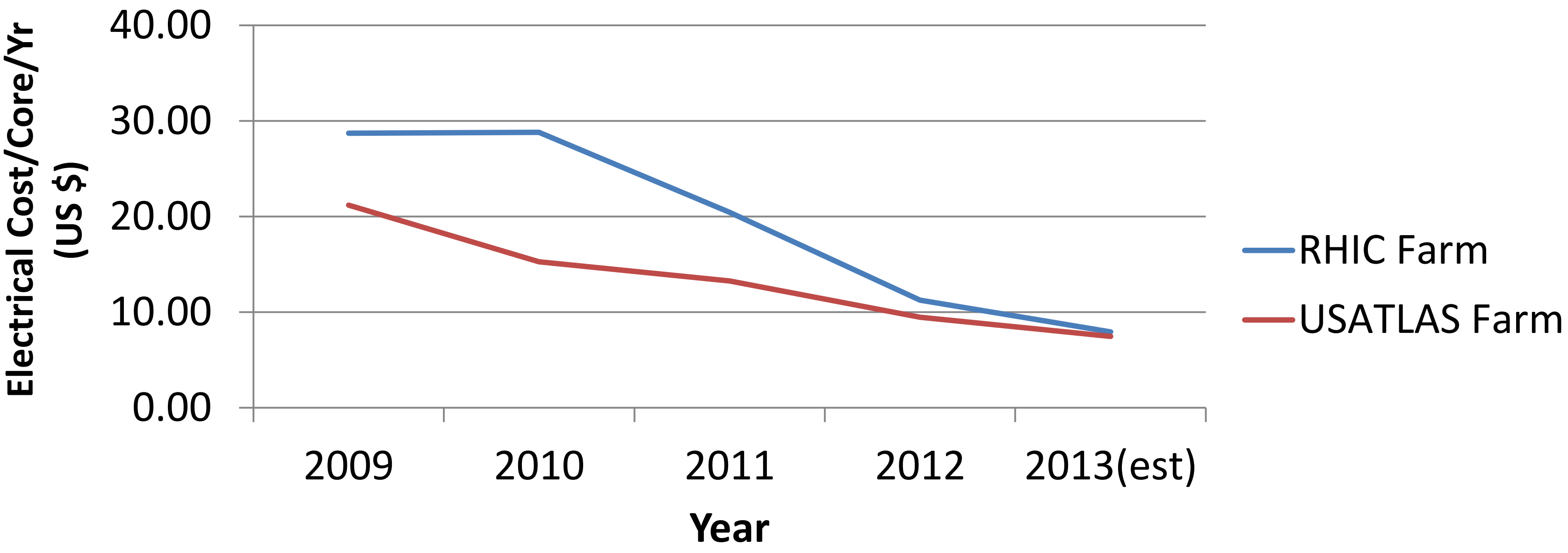
Network

- Network 1 GbE connectivity (switch, line cards, cabling, etc) for 900 hosts costs ~$450k
- Cost per core is ~$45 assuming our newest configuration (dual 8-core Sandybridge cpu’s)
- Assume network equipment is used during the lifetime of the Linux Farm hosts (4 years for USATLAS and 6 years for RHIC)
- Calculate cost per core per year by amortizing cost of network equipment over the lifetime of the servers

Software

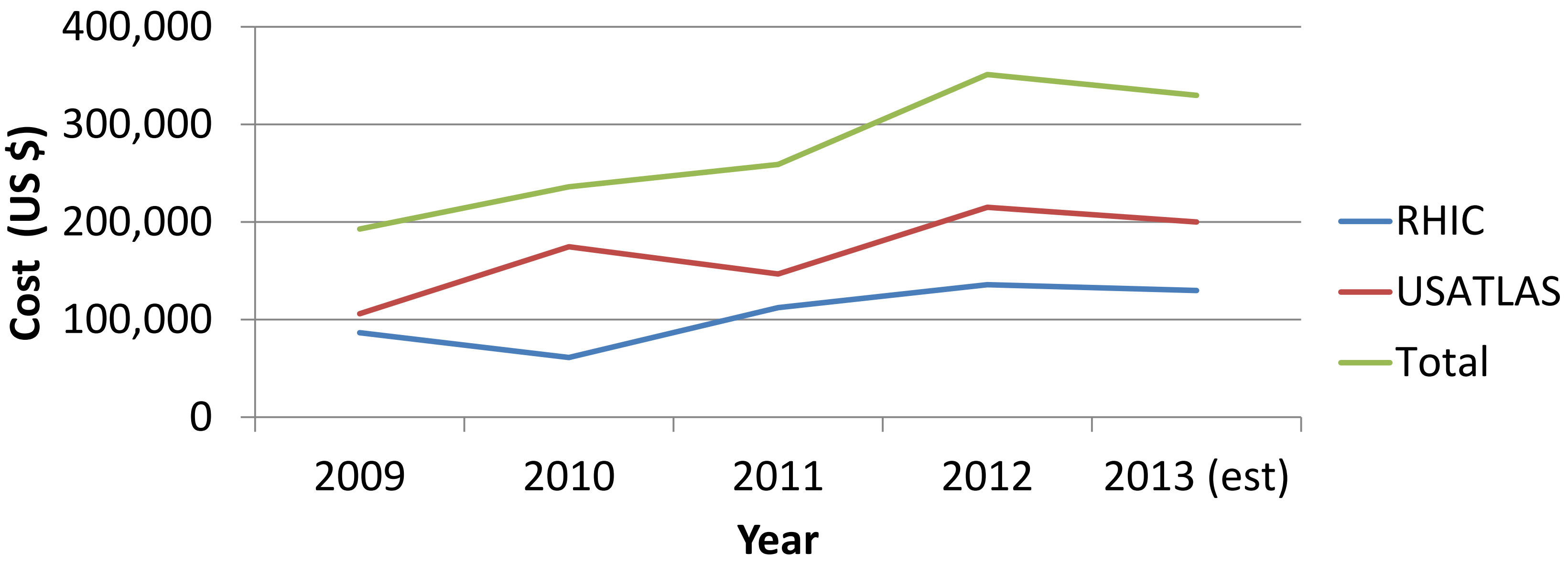
- Computing cluster uses mostly open-source software (free)
- Three exceptions
  - Ksplice (reboot-less kernel patching software)
  - Synapsense (real-time 3-D heat map of data center for environmental monitoring) – presentation by Alexandr Zaytsev at CHEP2013
  - Sensaphone (power monitoring and alert system used to detect when on UPS power)
- Negligibly small combined cost (~$3/core each for RHIC and USATLAS)

Electrical Costs



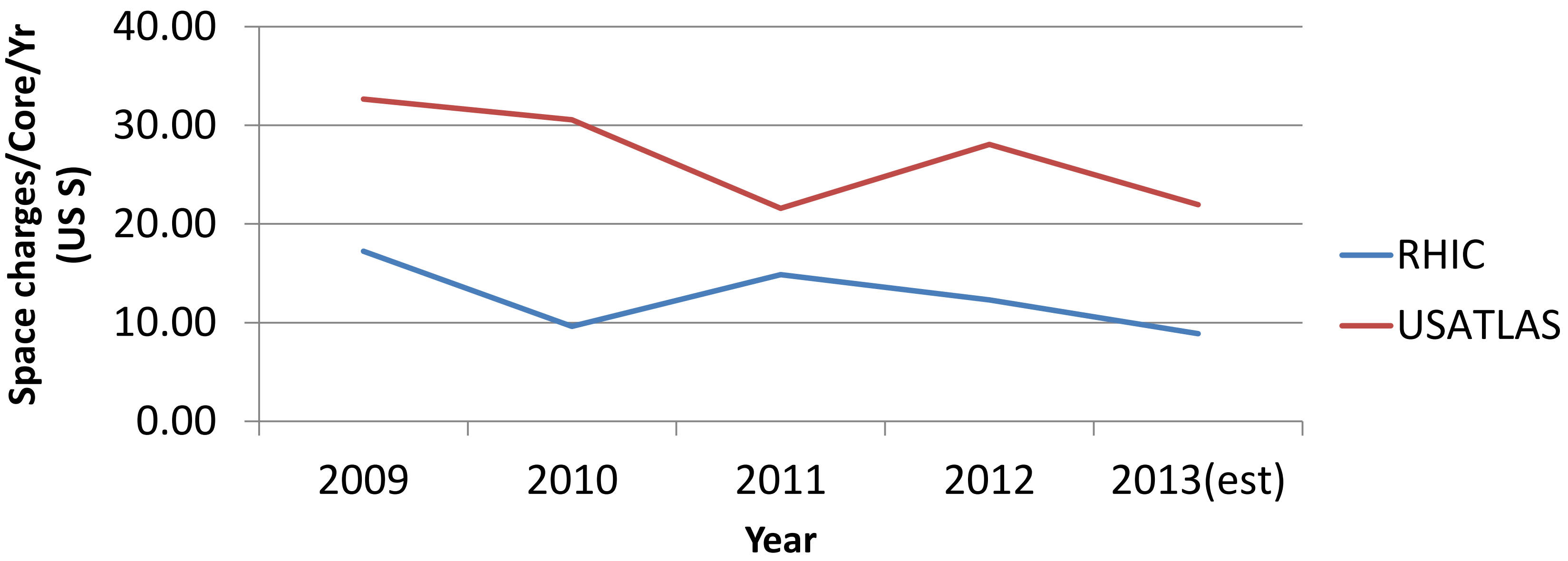
- Increasingly power-efficient hardware has decreased power consumption per core at the RACF in recent years
- RHIC costs higher than USATLAS due to differences in hardware configuration and usage patterns
- Average instantaneous power consumption per core was ~25 W in 2012

Overall Data Center Space Charges



- Overhead charged to program funds to pay for data center infrastructure (cooling, UPS, building lights, cleaning, physical security, repairs, etc) maintenance—upward trend a concern
- Based on footprint (~13,000 ft^2 or ~1200 m^2) and other factors
- USATLAS occupies ~60% of the total area.
- Rate reset on a yearly basis – not predictable

Space Charges for Servers



- Space charges incurred by Linux Farm clusters
- Estimated on approximate footprint occupied
- Charges/core dropping but overall footprint expanding

Historical Cost/Core

	USATLAS	RHIC
Server	$228/yr	$277/yr
Network	$28/yr	$26/yr
Software	$3/yr	$3/yr
Staff	$34/yr	$34/yr
Electrical	$12/yr	$16/yr
Space	$27/yr	$13/yr
Total	$332/yr ($0.038/hr)	$369/yr ($0.042/hr)

- Includes 2009-2013 data
- BNL-imposed overhead included
- Amortize server and network over 4 or 6 (USATLAS/RHIC) years and use only physical cores
- RACF Compute Cluster staffed by 4 FTE ($200k/FTE)
- About 25-31% contribution from other-than-server

Trends

- Hardware cost/core is flattening out
- Space charges are trending higher due to internal BNL dynamics
- Network cost trends in medium-term will be affected by technology choices
  - Remain at 1 GbE for now
  - Transition to 10 GbE in 2015?
    - $63/core (at 2013 prices) → substantially lower in 2 years?
  - Is Infiniband a realistic alternative?
    - Evaluation underway (estimate ~30-50% lower than 10 GbE)

Resiliency of Data Storage

- What if duplication of derived data at Tier1 is a requirement?
  - to avoid catastrophic loss of data (ie, fire, hurricane, etc)
  - to overcome hardware failure
- BNL is investigating:
  - Robotic tapes
  - NAS (BlueArc, DDN, MAPr, etc)
  - Worker nodes (via dCache, Hadoop, MAPr, etc)
  - Cloud (EC2, GCE, etc)

Cost of Data Duplication

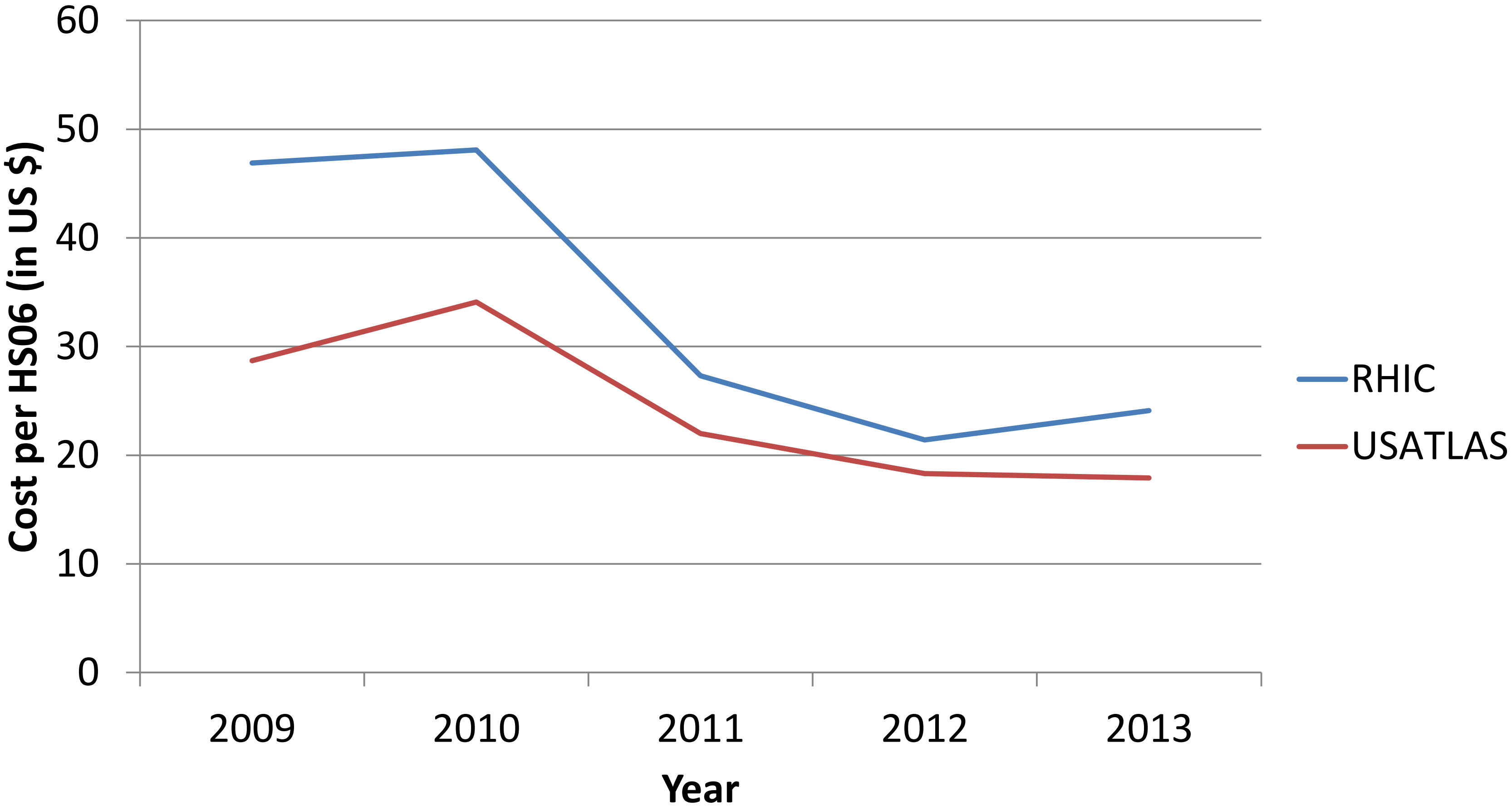
- Raw costs
  - Cloud (EC2) → $0.05/GB/month
  - Worker nodes → $0.08/GB
  - NAS → $0.20/GB
  - Tape (T10K) → $0.03/GB
- Example of 15 PB initial deployment on worker nodes
  - Current usage at RACF
  - Assume 100% duplication
  - Add $39/yr to cost/core of worker nodes
  - High-Availability requirement (UPS back-up costs?)

Summary

- Cost of computing/core at dedicated data centers compare favorably with cloud costs
  - $0.04/hr (RACF) vs. $0.12/hr (EC2)
  - Near-term trends
    - Hardware ↘
    - Infrastructure ↗
    - Staff ➡
    - Data duplication ↗
- Data duplication requirements will raise costs and complexity – not a free ride

Back-up slides

Cost per HS06



Additional RACF Details

- All RACF calculations assumed physical cores
  - In 2013, the RACF average physical core is rated at ~16 HS06
  - 2 ECU rating directly comparable to some EC2 instances
  - In practice, RACF hyperthreads all its physical cores, doubling the number of compute cores
- Hyperthreaded cores
  - adds ~30% to HS06 rating (~21 for 2013)
  - Approximately ~2.6 ECU