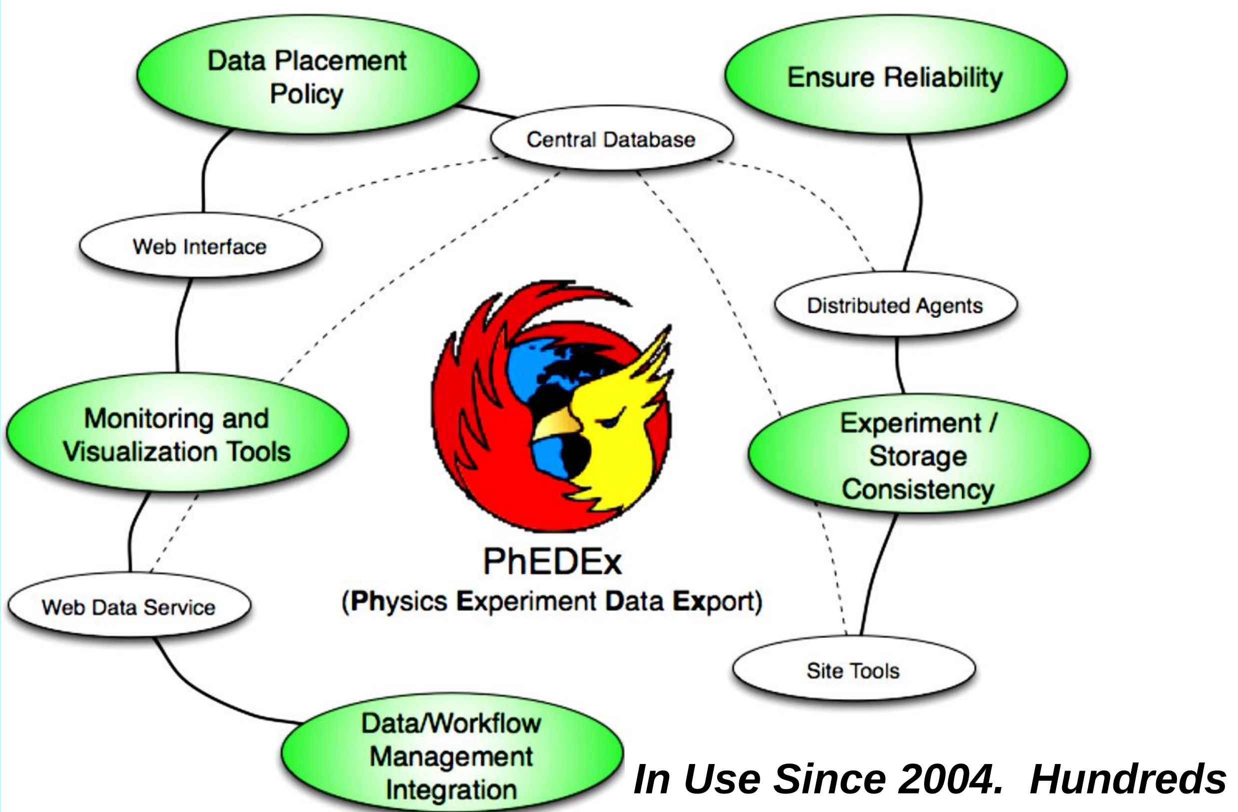
The LHC Computing Models and Concepts

Continue to Evolve Rapidly and Need To

How to Respond ?

PhEDEx (CMS): Physics Experiment Data Export https://cmsweb.cern.ch/phedex/



In Use Since 2004. Hundreds of Petabytes transferred

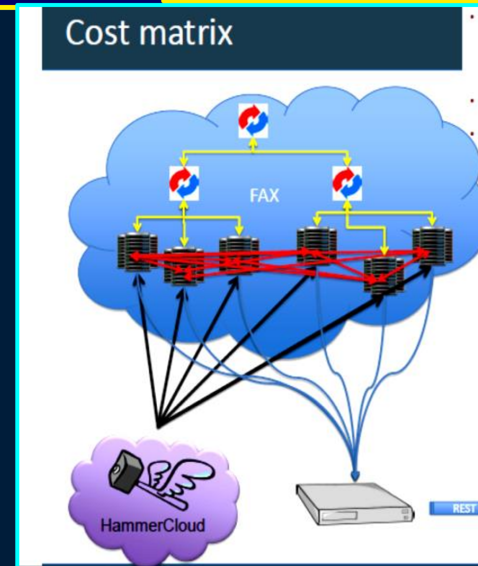
- Pull Model: Moderate the download rate to match the storage capabilities
- Tier2 sites for download still selected manually
- Download agents communicate via a Database ["blackboard"]
- Original assumption: network is a scarce and fragile resource
- Now need to adapt to higher speed + more reliable networks

- Desired Upgrades (Wildish): Faster switching among source-sites
- CMS-wide scheduling to avoid competition on shared links and end points
- Wider range of use cases: Possibly include downloads directly to desktops
- Dynamic Circuits to ensure, and control bandwidth

Integrating Network Awareness in ATLAS Distributed Computing

USE CASES

Kaushik De

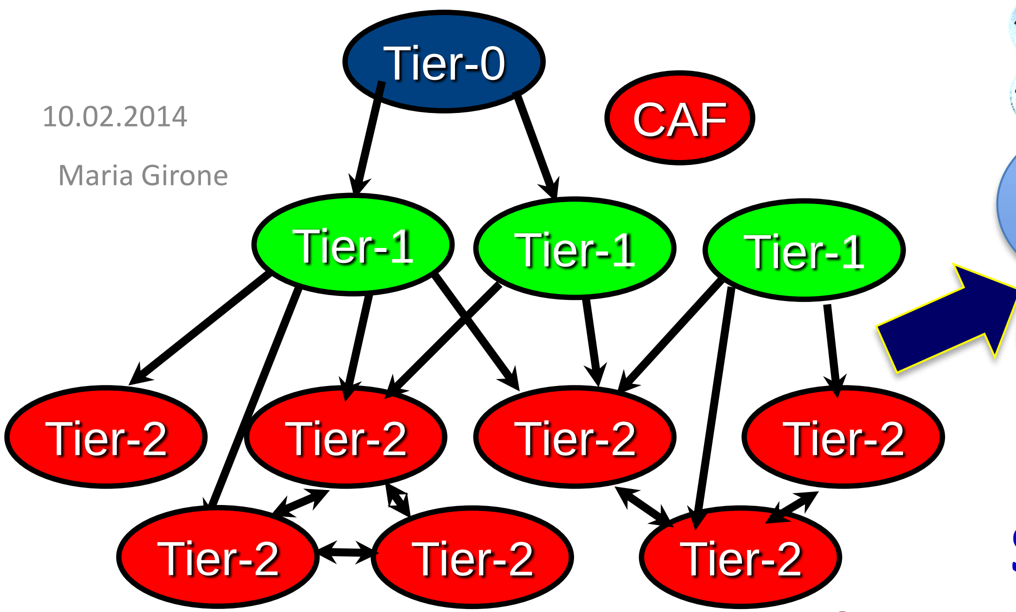


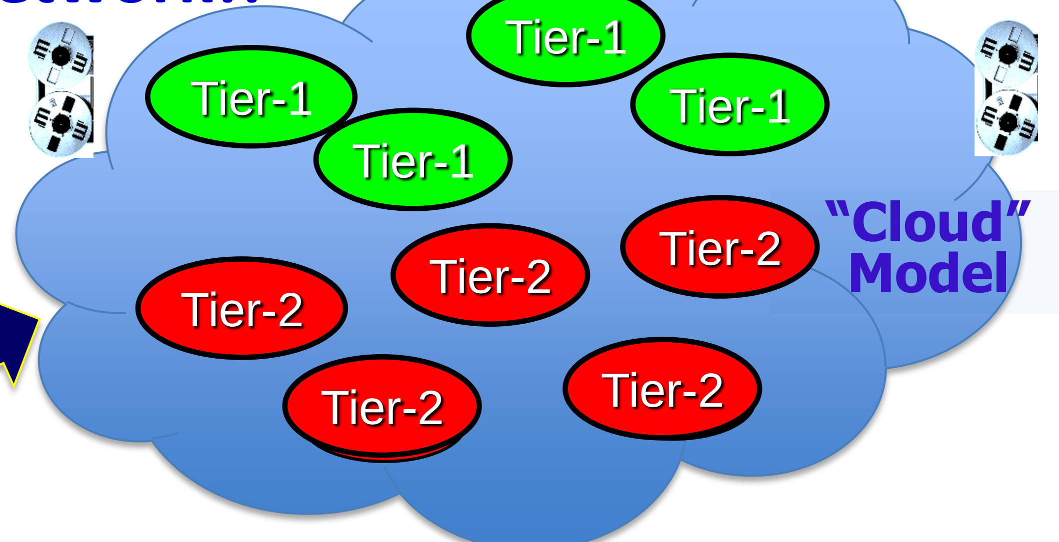
1. Faster User Analysis
   - Analysis jobs normally go to sites with local data: sometimes leads to long wait times due to queuing
   - Could use network information to assign work to 'nearby' sites with idle CPUs and good connectivity
2. Cloud Selection
   - Tier2s are connected to Tier1 "Clouds", manually by the ops team (may be attached to multiple Tier1s)
   - To be automated using network info: Algorithm under test
3. PD2P = PanDA Dynamic Data Placement: Asynchronous usage-based
   - Repeated use of data or Backlog in Processing → Make add'l copies
   - Rebrokerage of queues → New data locations

   PD2P is perfect for network integration
   - Use network for site selection – to be tested soon
   - Try SDN provisioning since this usually involves large datasets; requires some dedicated network capacity

CMS: Location Independent Access: Blurring the Boundaries Among Sites

- Once the archival functions are separated from the Tier-1 sites, the **functional difference between the Tier-1 and Tier-2 sites becomes small**
  - **Connections and functions of sites are defined by their capability, including the network!!**



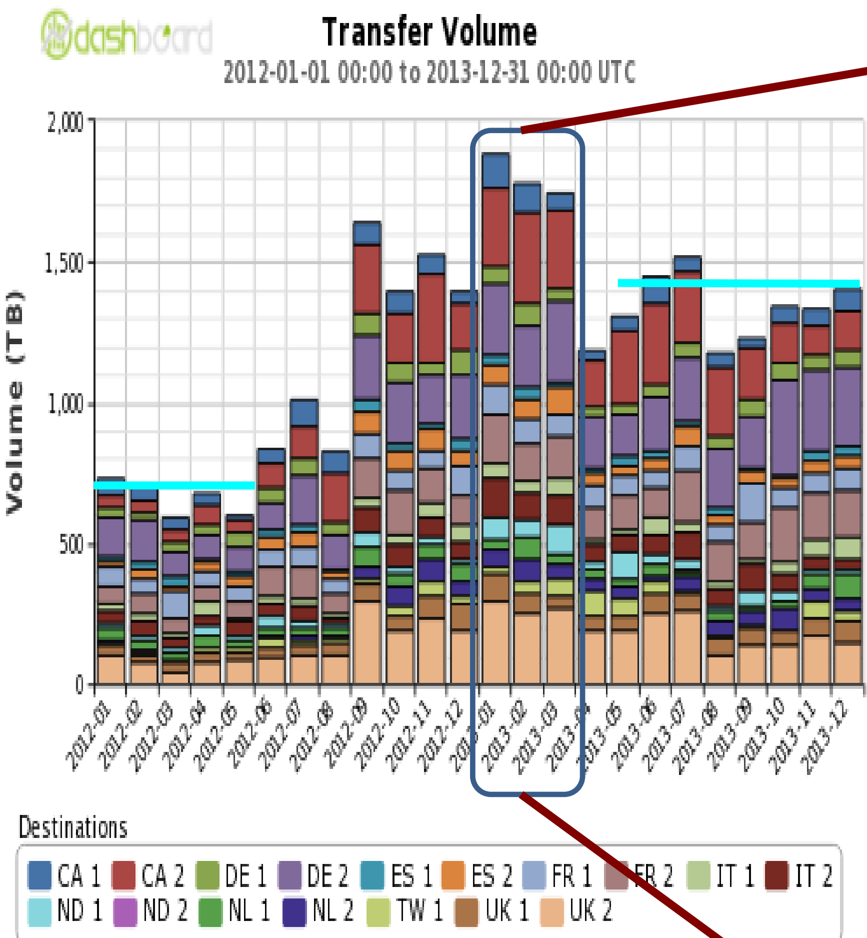


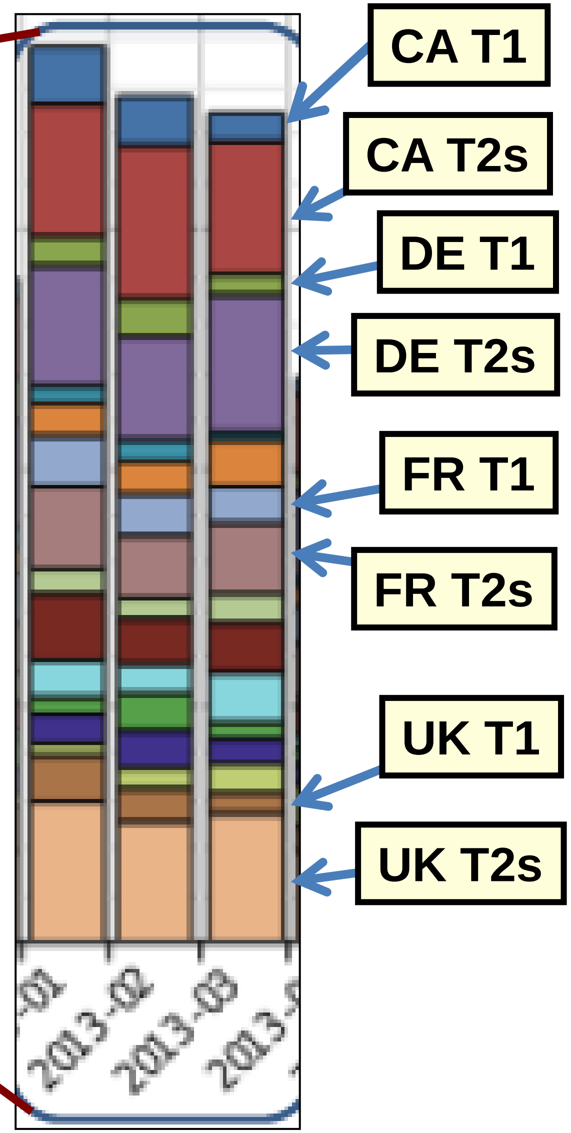
Scale tests ongoing:

Goal: 20% of data across wide area;

200k jobs/day, 60k files/day, O(100TB)/day

ATLAS: T1s vs. T2s from BNL
(2013 Winter Conference Preparations)





2H 2013 Volume was ~twice that of 1H 2012, even without data taking.

Exponential growth in data transfers continues, driven by Tier2 data usage.

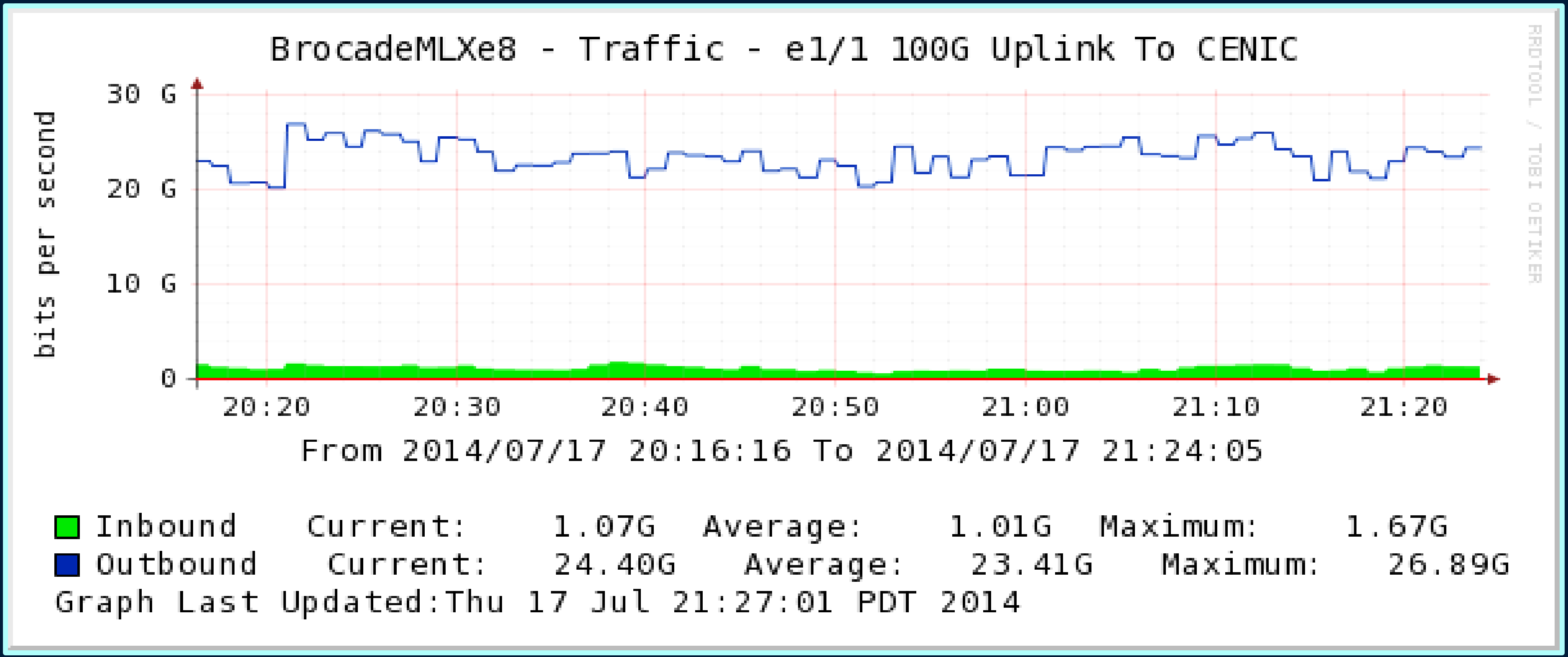
Expect new peaks by and during LHC Run 2

T2s in several regions are getting ~an order of magnitude more data from BNL than the associated T1s

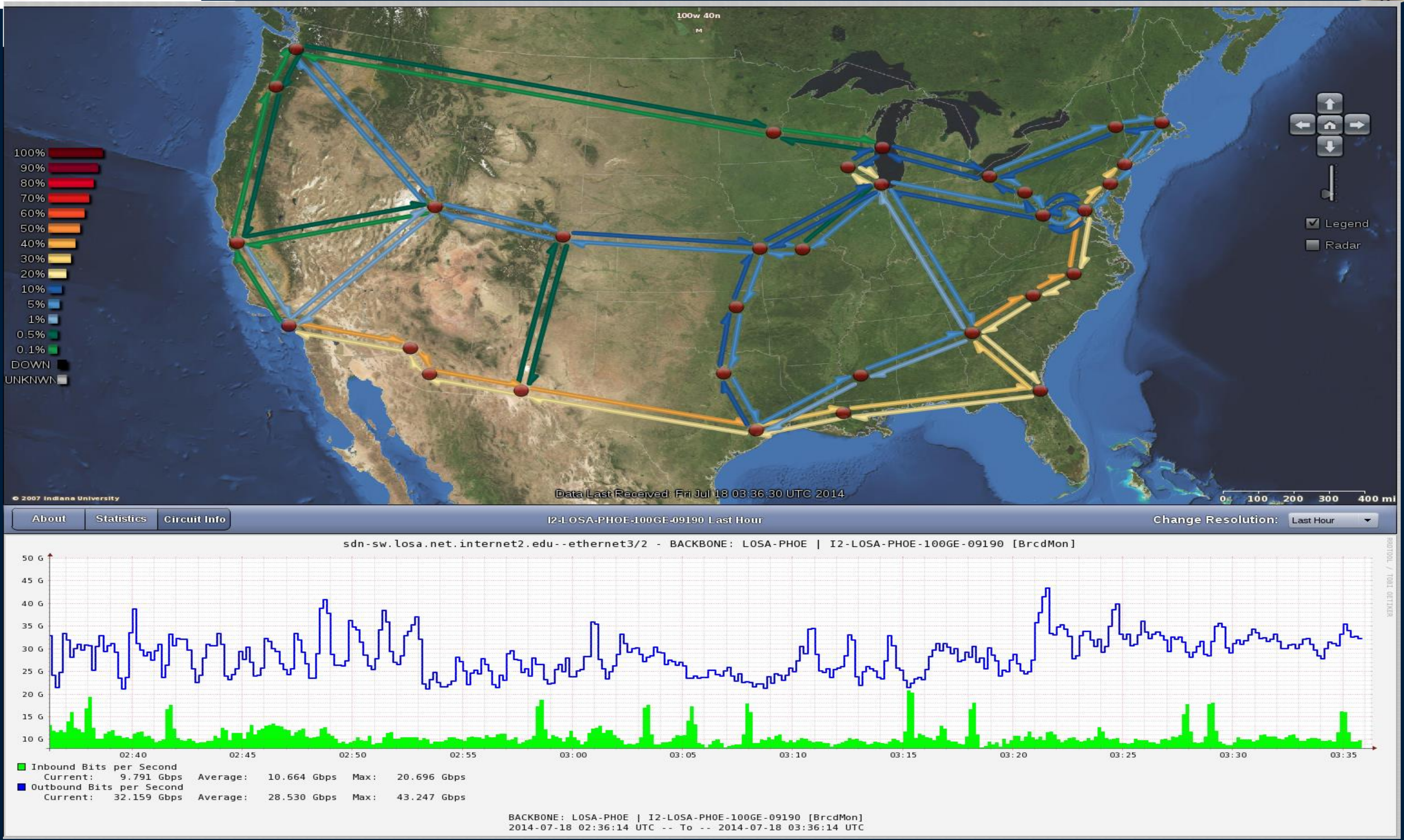
ANA-100 Link in Service July 16

Transfer Rates: Caltech Tier2 to Europe July 17

- Peak upload rate: 26.9 Gbps
- Average upload rate over 1h of manual transfer requests : 23.4 Gbps
- Average upload rate over 2h (1h manual+ 1h automatic) : 20.2 Gbps
- Peak rate to CNAF alone: 20 Gbps



Transfer Caltech ➡ Europe elevates usage of Internet2 to > 40% occupancy on some segments



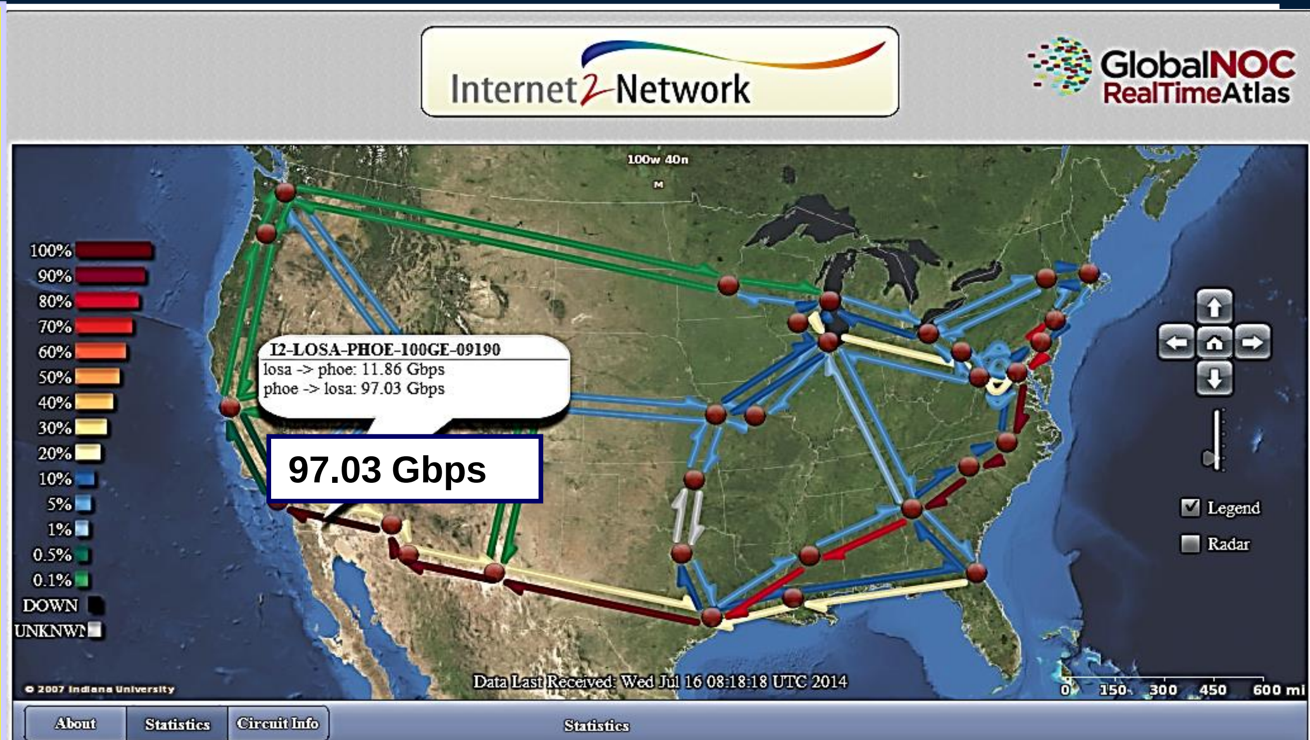
Internet2 Network Map

AL2S Traffic Statistics

Traffic peak 97.03 Gbps Phoenix - LA observed during these transfers

This is a possible limiting factor on the traffic received at Caltech

Microbursts are often not reported by the monitoring clients



Circuit Name	A -> Z	bits/sec		Packets/sec		Errors/sec	Z -> A	bits/sec		Packets/sec		Errors/sec
I2-RALE-WASH-100GE-08888	wash -> rale	82.56	Gbps	1.92	Mpps	0	rale -> wash	8.84	Gbps	1.15	Mpps	0
I2-ALBA-BOST-100GE-09210	bost -> alba	5.06	Gbps	526.89	Kpps	0	alba -> bost	5.5	Gbps	500.87	Kpps	0
I2-ASHB-WASH-100GE-09106	wash -> ash	2.92	Gbps	392.19	Kpps	0	ashb -> was	3.54	Gbps	432.59	Kpps	0
I2-ATLA-CHAR-100GE-07738	atla -> char	8.87	Gbps	1.15	Mpps	0	char -> atla	82.49	Gbps	1.92	Mpps	0
I2-ASHB-CHIC-100GE-11803	chic -> ashb	11.45	Gbps	1.02	Mpps	0	ashb -> chic	2.79	Gbps	406.25	Kpps	0
I2-ASHB-PITT-100GE-07737	ashb -> pitt	3.11	Gbps	350.85	Kpps	0	pitt -> ashb	2.1	Gbps	537.84	Kpps	0
I2-SEAT-SUNN-100GE-08997	seat -> port	77.79	Mbps	14.64	Kpps	0	port -> seat	27.63	Mbps	6391	pps	0
I2-CHIC-KANS-100GE-07745	chic -> kans	8.84	Gbps	647.05	Kpps	0	kans -> chic	7.03	Gbps	665.8	Kpps	0
I2-CHIC-COLU4-100GE-11554	chic -> colu	1.43	Gbps	202.3	Kpps	0	colu4 -> chi	3.41	Gbps	298	Kpps	0
I2-LOSA-SALT-100GE-07757	losa -> salt	909.48	Mbps	111.21	Kpps	0	salt -> losa	2.35	Gbps	149.16	Kpps	0
I2-PHIL-WASH-100GE-10867	wash -> phi	5.17	Gbps	722.9	Kpps	0	phil -> wash	74.47	Gbps	1.32	Mpps	0

Message: At anywhere near this level of capability, we need to control our network use, to prevent saturation as we move into production.

What Networks Need to Do

W. Johnston, ESnet Manager (2008)
On Circuit-Oriented Network Services

- For this essential approach to be successful in the long-term it must be routinely accessible to discipline scientists - without the continuous attention of computing and networking experts

- In order to
  - facilitate operation of multi-domain distributed systems
  - accommodate the projected growth in the use of the network
  - facilitate the changes in the types of traffic

➡ the architecture and services of the network must change

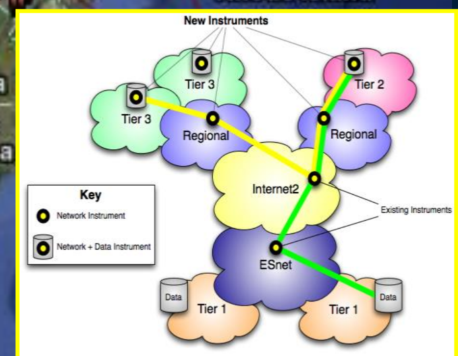
- **The general requirements for the new architecture are that it provide:**
  - **1) Support the high bandwidth data flows of large-scale science including scalable, reliable, and very high-speed network connectivity to end sites**
  - ✸ **2) Dynamically provision virtual circuits with guaranteed quality of service (e.g. for dedicated bandwidth and for traffic isolation)**
  - ✸ **3) provide users and applications with meaningful monitoring end-to-end (across multiple domains)**

Traffic Isolation; Security; Deadline Scheduling; High Utilization; Fairness

DYNES: Dynamic Circuits Nationwide System. Created by Caltech, Led by Internet2

DYNES goal is to extend circuit capabilities to ~50 US campuses Turns out to be nontrivial →	Partners: I2, Caltech, Michigan, Vanderbilt. Working with ESnet on dynamic circuit software

Extending the OSCARS scope; Transition: DRAGON to PSS, OESS



http://internet2.edu/dynes

Functionality will be an integral part of LHCONE point-to-point service: An Opportunity - Via SDN (OpenFlow and OpenDaylight)

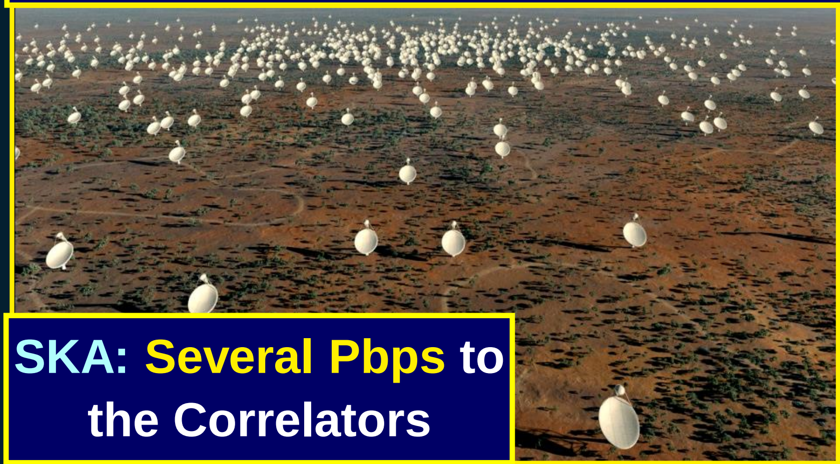
HEP Energy Frontier Computing

Decadal Retrospective and Outlook for 2020(+)

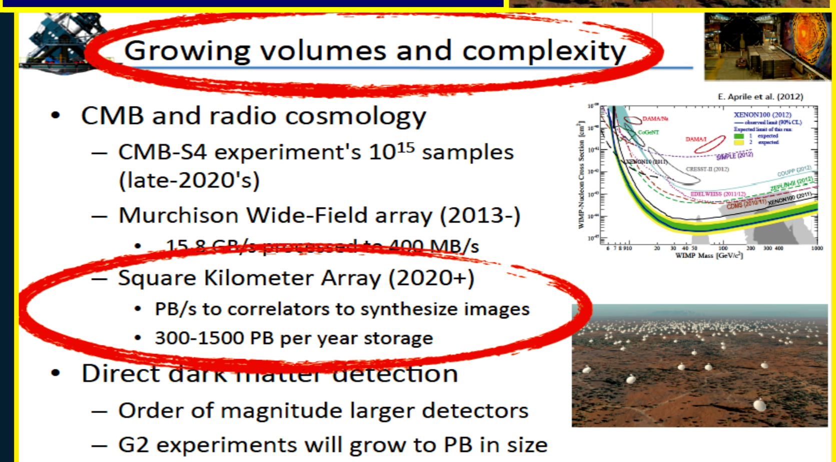
- Resources & Challenges Grow at Different Rates Compare Tevatron Vs LHC (2003-12)
  - Computing capacity/experiment: 30+ X
  - Storage capacity: 100-200 X
  - Data served per day: 400 X
  - WAN Capacity to Host Lab *100 X*
  - TA Network Transfers Per Day *100 X*
- Challenge: 100+ X the storage (tens of EB) unlikely to be affordable
- Need to better use the technology
  - An agile architecture exploiting globally distributed clouds, grids, specialized (e.g. GPU) & opportunistic resources
  - A Services System that provisions all of it, moves the data more flexibly and dynamically, and behaves coherently;
- Co-scheduling network, CPU and storage

Snowmass Computing Frontier Sessions

Challenges Shared by Sky Survey, Dark Matter and CMB Experiments.
SKA: 300 – 1500 Petabyes per Year



SKA: Several Pbps to the Correlators



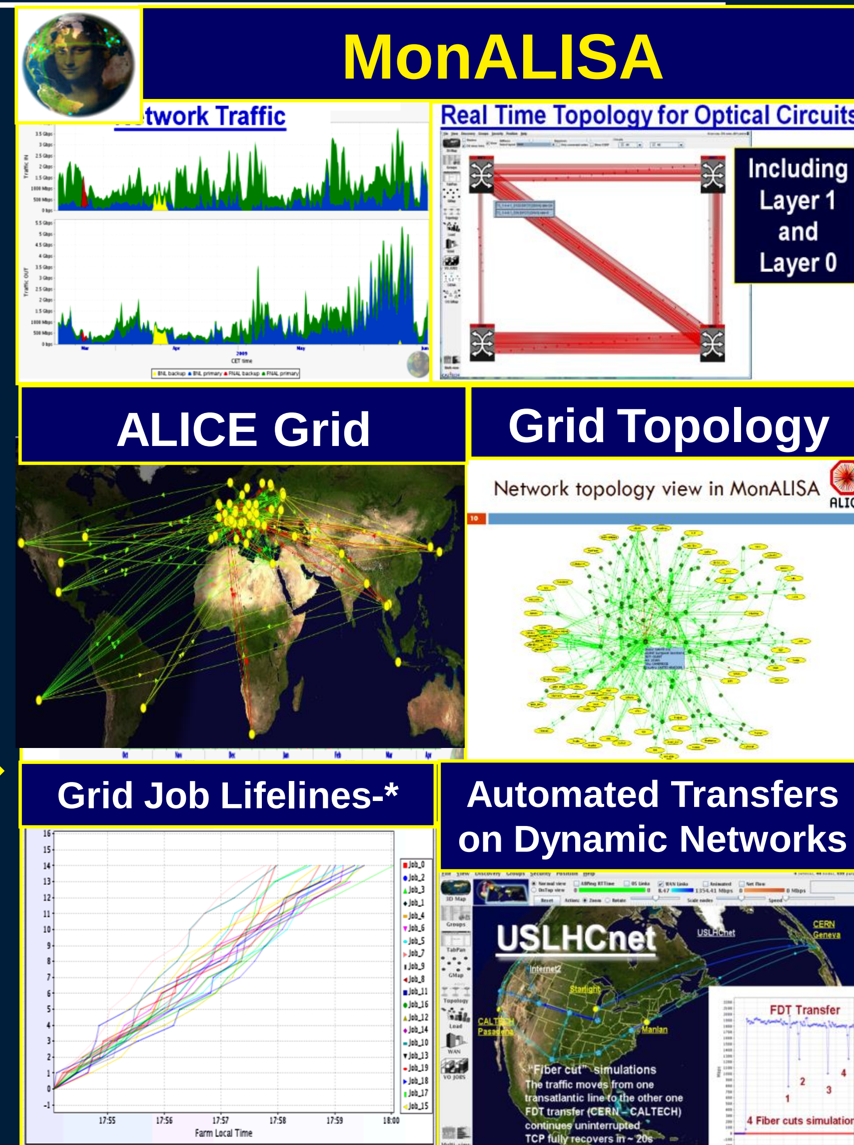
Key Issue and Approach to a Solution:

Next Generation System for Data Intensive Research

- Present Solutions will not scale
- We need: an agile architecture exploiting globally distributed **grid, cloud, specialized (e.g. GPU) & opportunistic computing resources**
- A Services System that **moves the data flexibly and dynamically, and behaves coherently**
- **Examples do exist**, with smaller but still very large scope

MonALISA ⟶

- A pervasive, agile autonomous agent architecture **that deals with complexity**
- **Developed by talented system developers** with a deep appreciation of networks



Message on Science Drivers
Discoveries and Global Impact

- Reaching for the next level of knowledge - New "*Invariants*": (1) Data + Network Intensiveness (2) Global Reach
- Instruments with unprecedented reach (energy, intensity; speed & scope of investigations; dealing with complexity; precision)
- Mission Focus: Envisage and develop the solutions (physics); Design and build new detectors, instruments, methods; Design and build the Systems
- Persistence: Program Lasting Years → Ph. D Units → Decades
- The Imperative of *New Models: vastly* greater operational efficiency leads to greater science output … and *Discoveries*
- Key Questions for the Scientist/Designer/Discoverer:
  - How to Solve the Problems; How to bring the Discoveries in Reach
  - Grappling with many fields of science, and many fields of technology
- The Art and Science of (Network) Requirements:
  - Bottom Up: Fixed budgets;
  - Top Down - Evolution and Revolution – to the next generation
  - Tracking emerging technologies, capabilities, affordability
  - Asking the Right Question: Maximizing capability within the budget