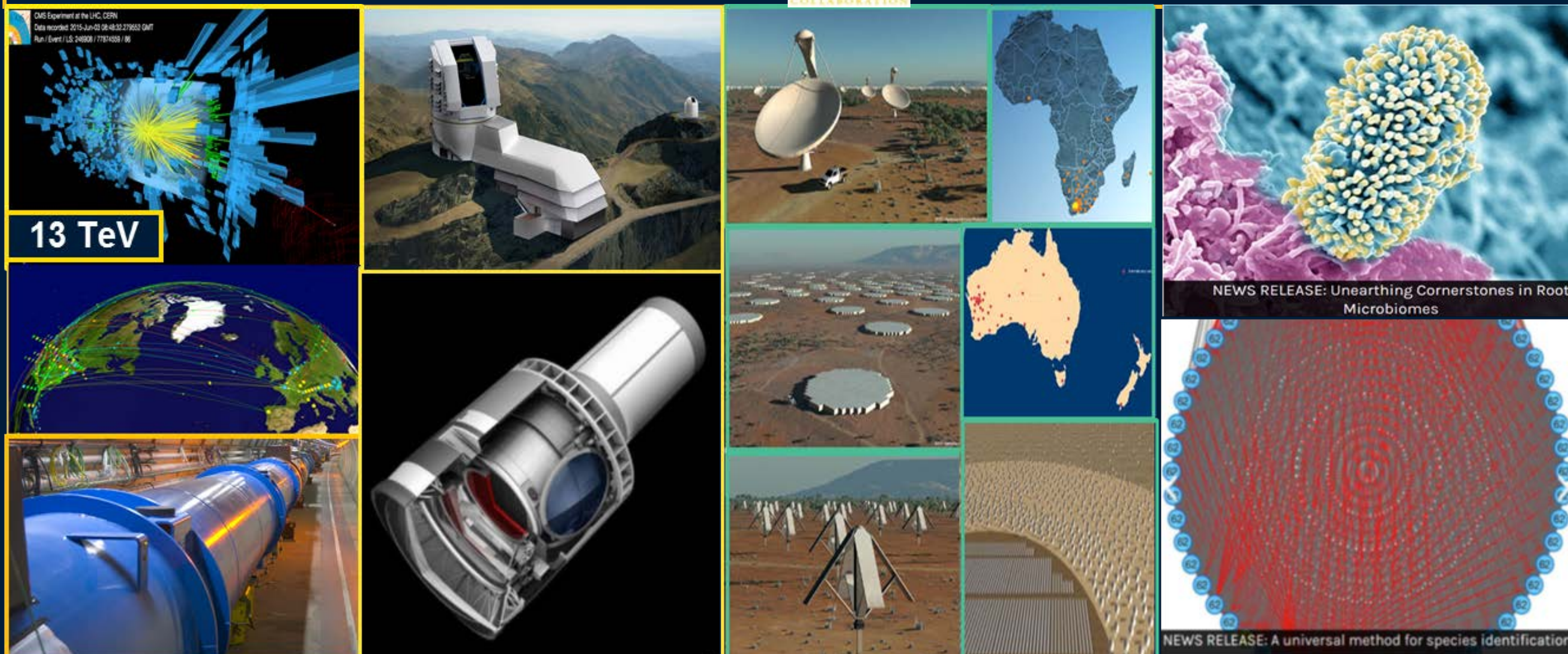
Next Generation Networks and Systems for Data Intensive Science



LHC | LSST | SKA | Joint Genome Institute

- *LHC Run1: Discovery of a New Boson*
- *LHC Run2+: Beyond the Standard Model*
- *Data Intensive Exascale LCFs and SDN EcoSystems*

Gateways to a New Era

Harvey B Newman, Caltech

Network Research Problems + Challenges for DOE Scientists Workshop

Bethesda, February 1 2016

Entering a New Era of Technical Challenges as we Move to Exascale Data and Computing

- The largest science datasets today, from LHC Run1, are 300 petabytes
  - Exabyte datasets are on the horizon, **by the end of Run2 in 2018**
  - These datasets are foreseen to grow by another 100X, to the ~50-100 Exabyte range, **during the HL LHC era from 2025**
- The reliance on high performance networks will thus continue to grow **as many Exabytes of data are distributed, processed and analyzed at hundreds of sites around the world.**
- **As the needs of other fields continue to grow,** HEP will face increasingly stiff competition for the use of large but limited network resources.

1 EB = 2 milligrams of DNA

Earth Observation

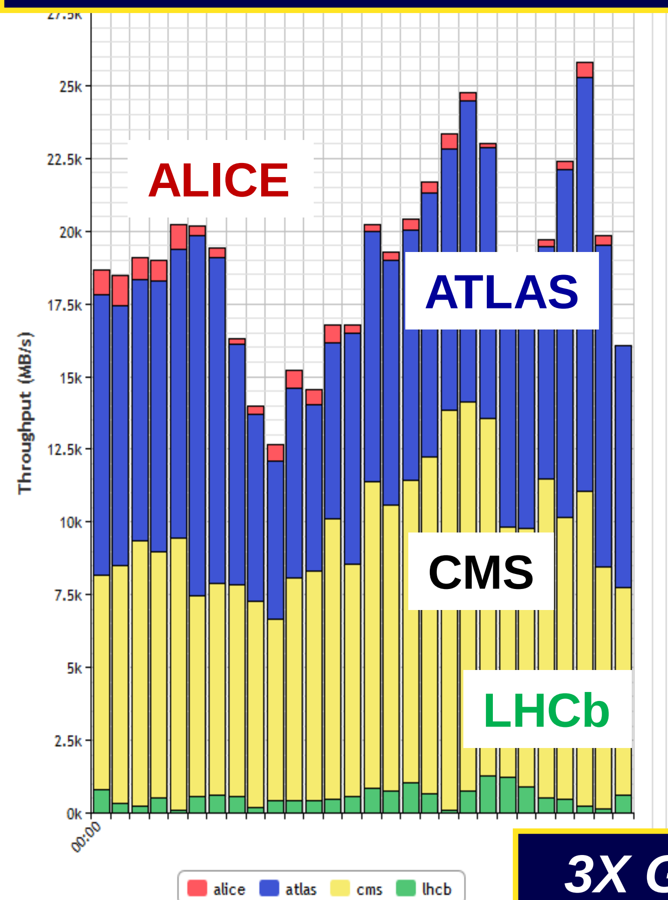
A New Era of Exploration and Discovery in Data Intensive Sciences: Challenges

- Scale of Datasets and Network Traffic [LHC view]
  - Data Stored: from 300 PBytes now to ~1 Exabyte by 2019
  - Traffic Flows: from 20-50 Gbps (with peaks to 90 Gbps) now; More sites and diverse paths; 100G+ flows “when possible”
  - Aggregate Transfers (WLCG): 50 PBbytes/month in Fall 2016; Projects to 1 Exabyte/month by ~2020 to 10 EB/Month by ~2024
- Complexity: Of diverse workflows, global topology, flow paths
  - Workflow: Organized dataset transfers, job output files, object collection access
  - Matching Jobs to Data: Redirection of data as needed
  - Of global topology (peerings and interconnections): LHCONE example
- Reactive, sometimes chaotic operations
  - Lack of network awareness by applications and users
  - Highly varied level of network capability among sites and regions
  - Lack of monitoring, interaction, feedback; moderation of user behavior
- Drivers: Bottom Up and Top Down Combined
  - Mission Need: LHC to LSST to Genomics; the Exascale Imperative
  - Technology and Opportunity Drivers: Low cost servers (CPU, storage) + apps capable of 100-500 Gbps with 100G NIC(s) and FDT; new memory, storage, OS

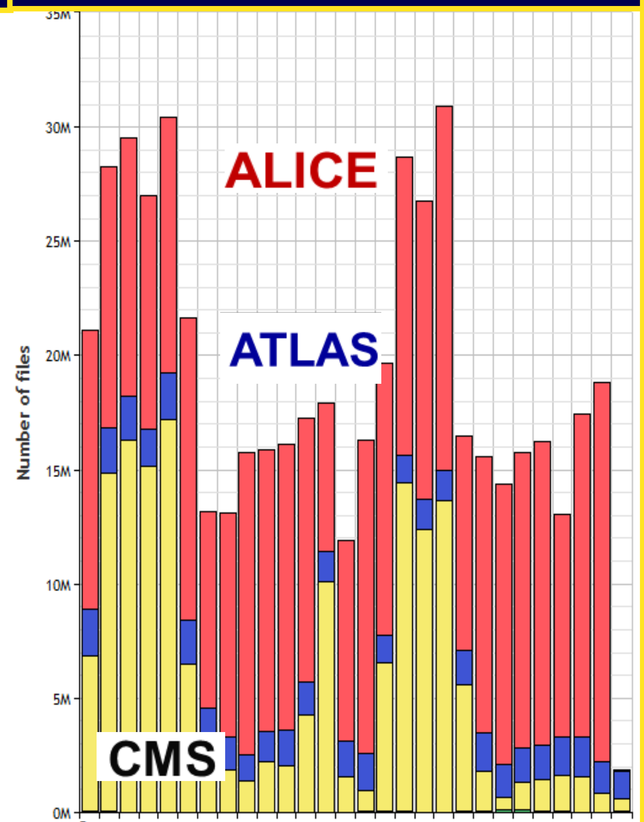
Complex Workflow: the Flow Patterns Have Increased in Scale and Complexity, even at the start of LHC Run2

WLCG: 170 Centers in 40 Countries. 2 Million Jobs Per Day

Transfer Throughput



Transfers Done/Day



3X Growth March – October

20 GBytes/s Typical

To 35 GBytes/s Peak Transfer Rates

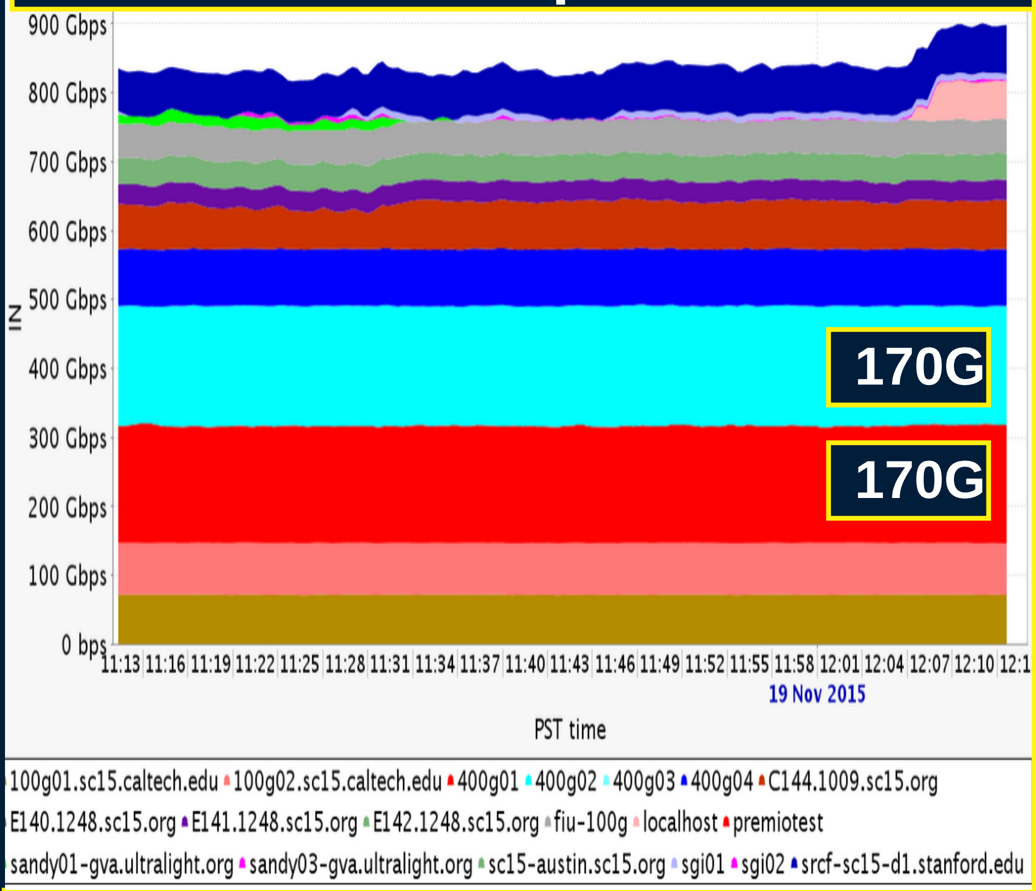
Complex Workflow

- Multi-TByte Dataset Transfers
- Transfers of 12-31 Million Job Output Files Daily
- Access to Tens of Millions of Object Collections/Day
- >100k of remote connections (e.g. AAA) simultaneously

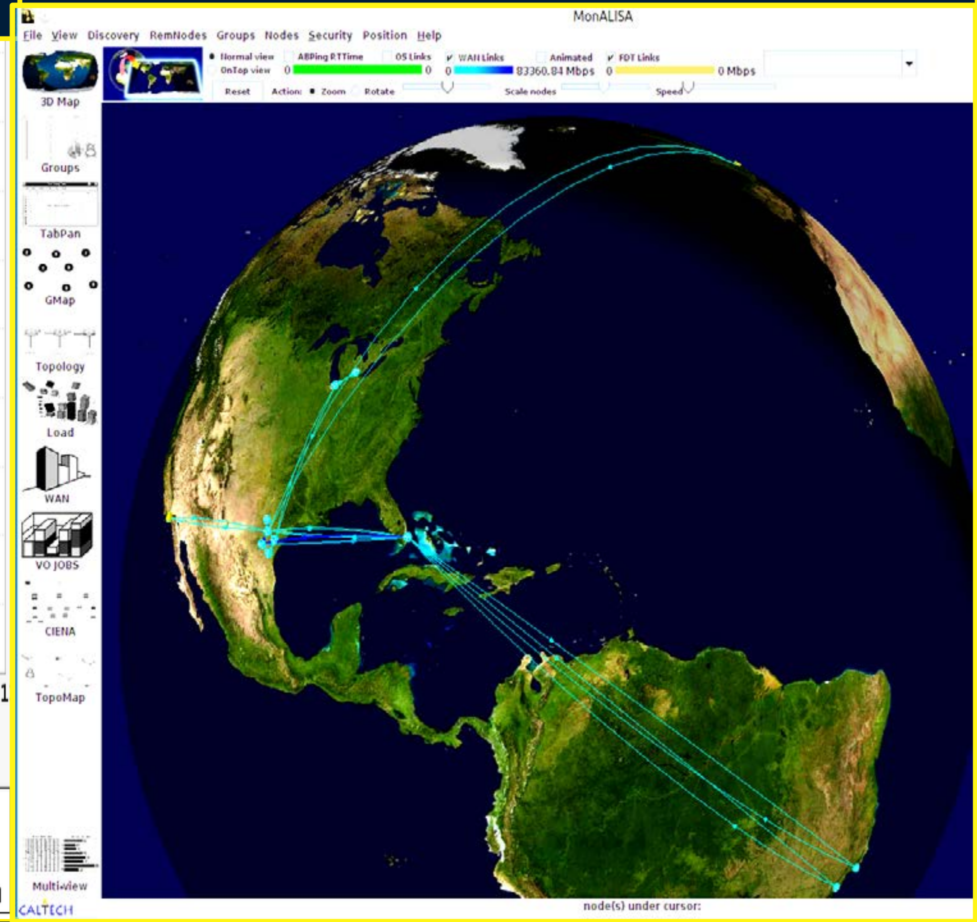
WLCG Dashboard Snapshot Sept-Oct. Patterns Vary by Experiment

Caltech and Partners Terabit/sec SDN Driven Agile Network: Aggregate Results

900 Gbps Total
Peak of 360 Gbps in the WAN



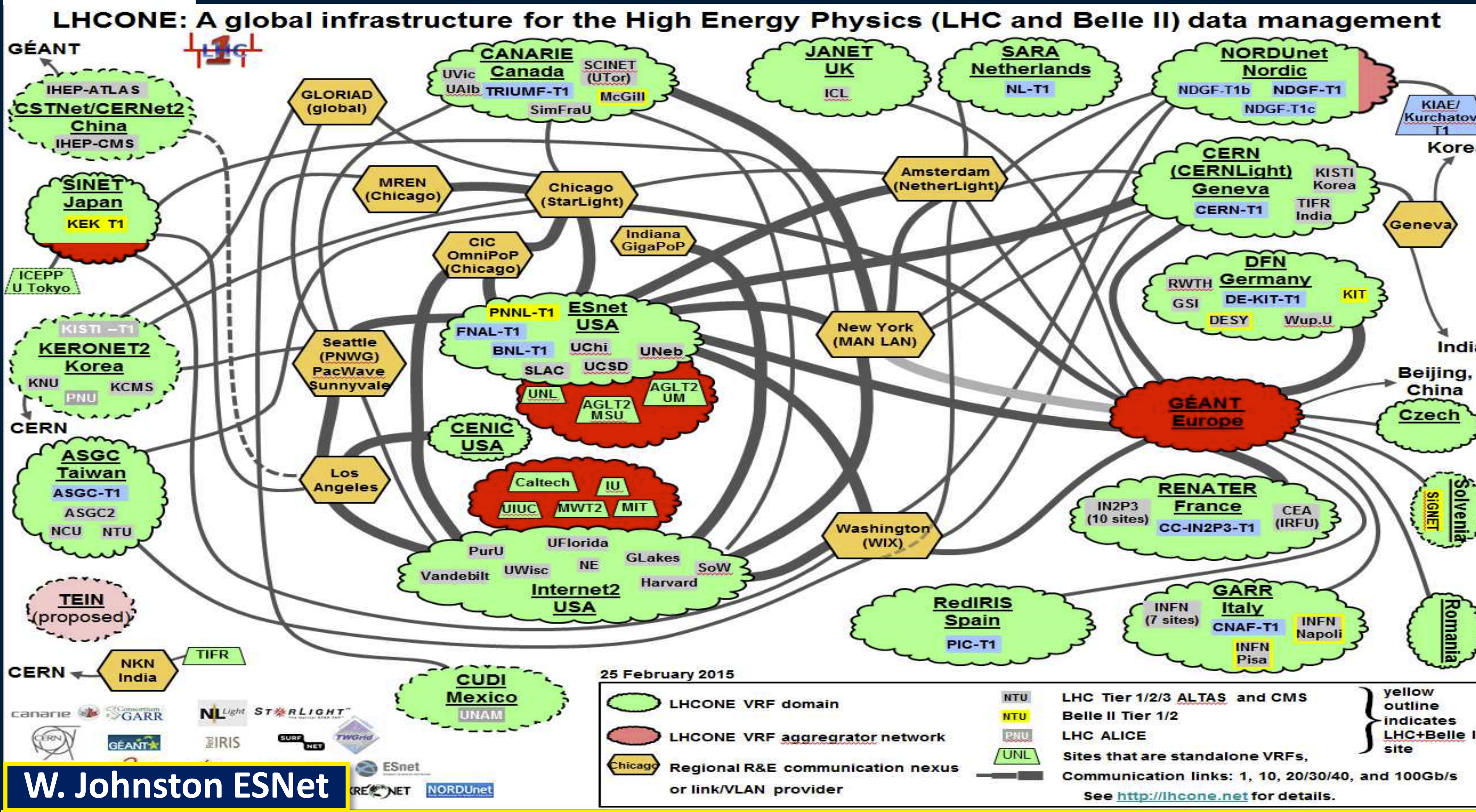
MonALISA Global Topology



Single port smooth Flows up to 170G; *120G over the WAN.* With Caltech's FDT TCP Application http://monalisa.caltech.edu/FDT

LHCONE: a Virtual Routing and Forwarding (VRF) Fabric

A global infrstructure for HEP (LHC and Belle II) data management



W. Johnston ESNet

The Major R&E Networks Have Mobilized on behalf of HEP

A stepping stone to our joint future: for data intensive sciences

The LHC: Spectacular Performance

Data Complexity: The Challenge of Pileup

~3.5 X 10^{15} pp Collisions
1M Higgs Bosons created in Run 1

~50 Vertices, 14 Jets, 2 TeV

Run2 and Beyond will bring:

- Higher energy and intensity
- Greater science opportunity
- Greater data volume & complexity
- A new Realm of Challenges

Average Pileup

Run 1 21
Run 2 42
Run 3 53
HL LHC 140-200

Achieving a New Era of Exploration and Discovery

Rising Challenges and Needs

- We are midway in the 7-8 year cycle of the present 100G network generation
  - It is getting too easy to match the capacity of production networks today, with 1000s of compute nodes, or with a very few well configured DTNs
  - Exhaustion of network resources may come before the next generation
- The intensity of usage will increase as the LHC program progresses; The outlook is for increasingly chaotic operations unless:
- Network awareness of users and applications is raised
- Interaction and feedback among user applications and network operations is implemented: getting users and the network "on the same side"
  - Service Classes including preferred service for those that plan, interact and well-use allocated resources
- Greater Predictability (of transfers in progress, scheduled and planned) is achieved, through intelligent network services and pervasive monitoring
- Greater intelligence and agility is implemented in the network:
  - Short term: path selection, flow steering, load balancing, allocation of scarce resources; strategic rebalancing
  - Longer term: managing resource allocations, identifying "reliable" requestors, fair-sharing over week/month/year

✸ Bottom Line: A real-time end-to-end system with a top down view, pervasive monitoring + a management paradigm (goals, strategy and tactics) is needed

Use of high capacity reliable networks opens up the “phase space” of available and affordable resources

10,000 feet overview

O.Gutsche’s talk at CHEP 2015

Grid

- Virtual Organizations (VOs) of users trusted by Grid sites
- VOs get allocations ➜ **Pledges**
  - Unused allocations: opportunistic resources

Cloud

- Community Clouds - Similar trust federation to Grids
- Commercial Clouds - **Pay-As-You-Go** model
  - Strongly accounted
  - Near-infinite capacity ➜ **Elasticity**
  - Spot price market

HPC

- Researchers granted access to HPC installations
- Peer review committees award **Allocations**
  - Awards model designed for individual PIs rather than large collaborations

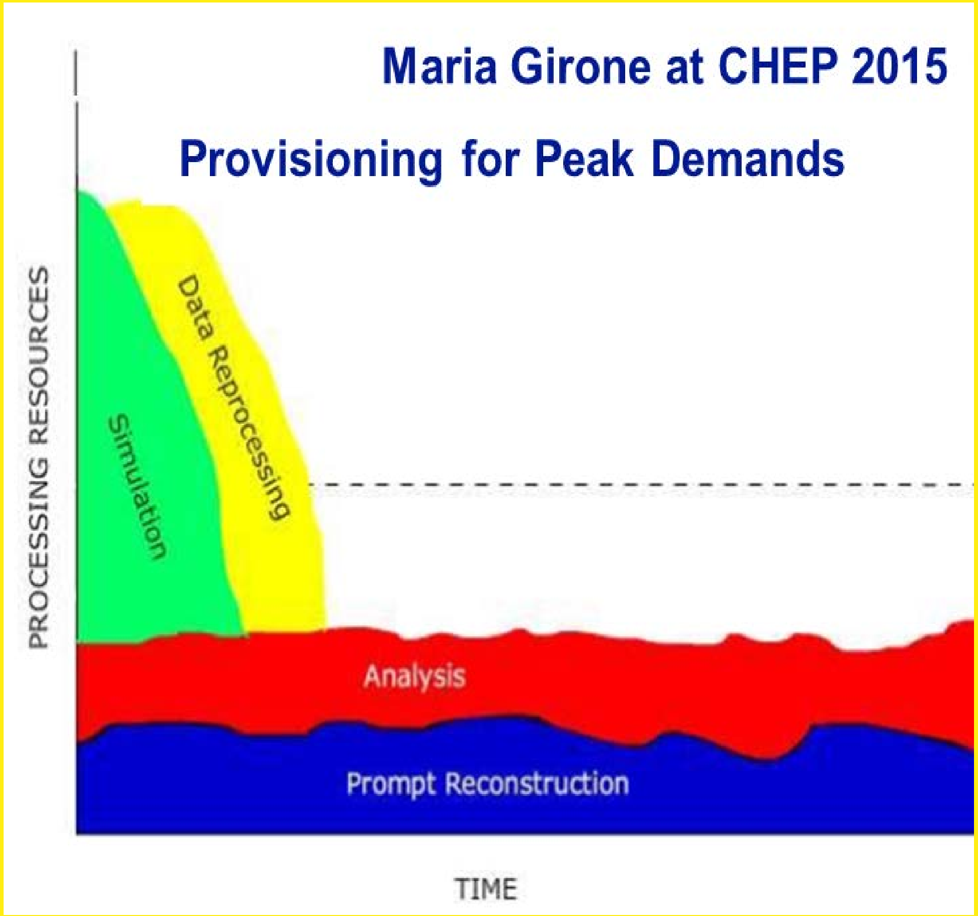
Grant Allocation

Provisioning for Peak Demands

Vision of short turn-around times for HEP major workflows

- Short latencies in particular in analysis workflows are important for science efficiency
- Using resources from a larger pool when they are needed, should also result in more cost-effective solutions

- Separating the processing and storage services allows them to scale independently
- e.g. ATLAS and CMS are looking at ways to double available resources for periods of time, Using Amazon services
  - CMS: 56k Core MC production now underway



Provisioning for peak requires that we use pooled resources

➡ Clouds and/or large HPC Centers!

Achieving a New Era of Exploration and Discovery

Elements for Success

- Bringing basic capability to the community: Beyond "best practices"
  - Workshops and field deployments of well configured DTNs + applications
- Reducing heterogeneity: bringing all areas to a minimum level
  - Identification and resolution of problem sites, links, regions
- Raising network awareness
  - A paradigm of "interaction leads to improved service"; non-interacting applications/users get best effort service, mapped onto a limited fraction of the available network resources
  - Interaction is imperative for users/groups that have a major impact on the networks and/or require priority service on demand
- Development of new "real time systems, driven by application / site/ network interactions, with true end-to-end operations
  - Agent based architectures with great resilience and adaptability
  - Monitoring systems with great scalabilility, pervasiveness, MTBF
- SDN is a natural pathway
  - Intent-based networking will ease the task for some users; but greater transparency implies greater intelligence "under the hood"
  - Which SDN: ODL, ONOS; OpenvSwitch, Openstack, or other ?
  - A powerful, rapidly advancing direction: but highly diverse and fluid

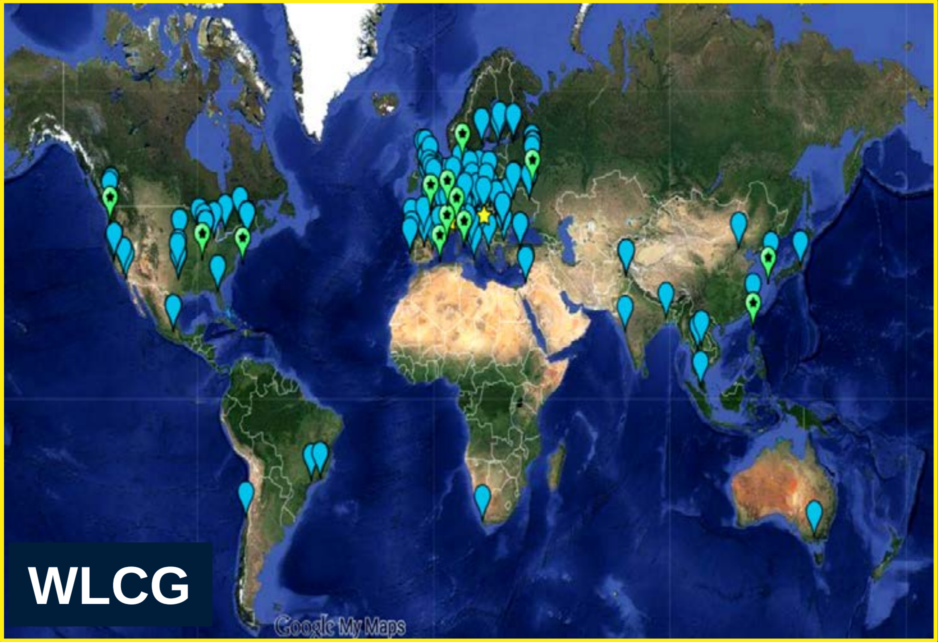
Achieving a New Era of Exploration and Discovery
Concepts and Issues

- System Architectural Concepts: Open systems with simple characteristics versus more intelligent, deterministic, predictable systems that are internally more complex, including stateful "end to end system services". Examples
  - Emerging network operating systems to manage network/site/user interactions based on intents
  - Real-time distributed systems: technically possible but difficult outside the single project "domain"
  - Monitoring systems that can track end-to-end operations: require sufficient data access across multiple sites and domains
  - Information Centric Networks: how much state is needed/wanted for data discovery, caching & routing as a function of data transaction size
- Network system/user interaction models: trading engagement and rule-based behavior for resources beyond the lowest common denominator
- Choices: diversity versus emerging standards and the ability to build on a common base. Choices that are more than technical survival of the fittest.
- Resource Sharing: Mission oriented tasks versus general service; degree of mission orientation and hence the resource allocation profile: varying by network, region and domain
- Consistent operations: high water marks for individual + aggregate large flows

Vision: Next Gen Integrated Systems for Exascale Science: Synergy ➡ a Major Opportunity

Exploit the Synergy among:

1. Global operations data and workflow management systems developed by HEP programs, **being geared to work with *increasingly diverse and elastic resources* *to respond to peak demands***
   - **Enabled by** distributed operations and security infrastructures
   - **Riding on** high capacity (but mostly still-passive) networks



2. Deeply programmable, agile software-defined networks (SDN) Emerging as multi-domain network "operating systems"

   + New network paradigms focusing on content: from CDN to NDN

3. Machine Learning, modeling and simulation, and game theory methods **Extract key variables; optimize; move to** real-time self-optimizing workflows

✸ The Watershed: A new ecosystem with ECFs **as focal points in the global workflow;** meeting otherwise daunting CPU needs

Achieving a New Era of Exploration and Discovery
Mechanisms and Choices: What is the Role ?

- **Dynamic Circuits: How hard or flexible the bandwidth guarantees ? How dynamic in time and capacity ?**
- **Role of Slices: Who can (qausi-permanently) reserve a slice ?**
- **Flow Steering: Classes of work definition and parameters; Authorization, priority; Dynamism: How often and how extensive ?**
- **Load balancing: Tactical and strategic; Dynamism questions as above**
- **Protocols: Policies and operations on inefficient or unfriendly protocols, and "inefficient" users; Protecting a valuable resource**
- **Layer 1 as well as Layer 2: Where and when**
- **For several of the above: guidelines on agility versus stability**
- **Coexistence of heterogeneous domains: with varying architecture, topology, technologies, performance, and Policies**
- **New "stateful" models of use and sharing: "Cost" based, quota based, role/priority based; top level metric based**
- **Effective Metrics: Throughput, resource usage, average + maximum time to completion, overall user-organization "satisfaction". *What* is optimal ?**

THANK YOU!

Harvey Newman

newman@hep.caltech.edu

ADDITIONAL ILLUSTRATIVE SLIDES FOLLOW

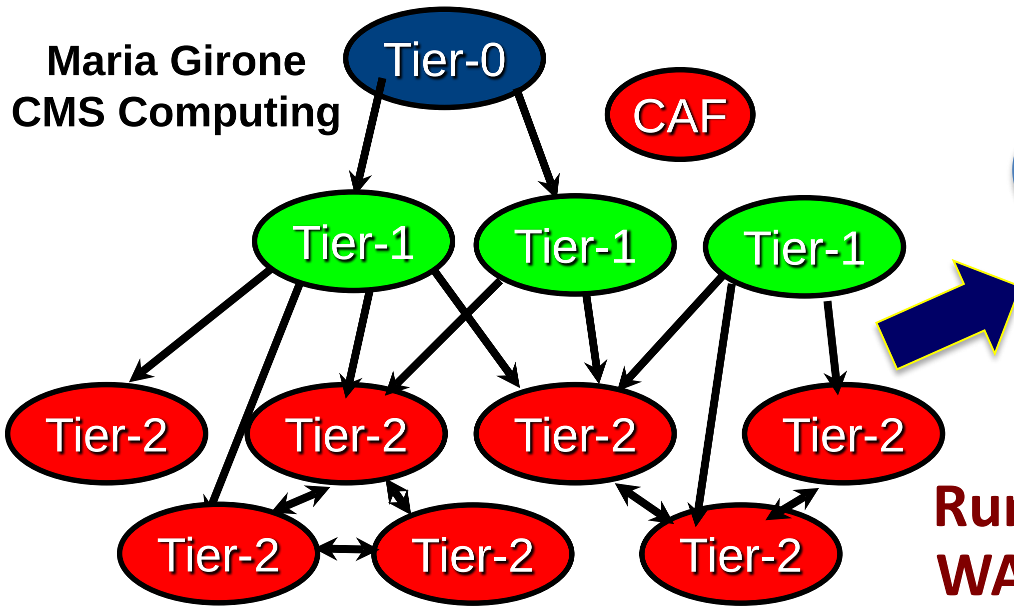
Harvey Newman

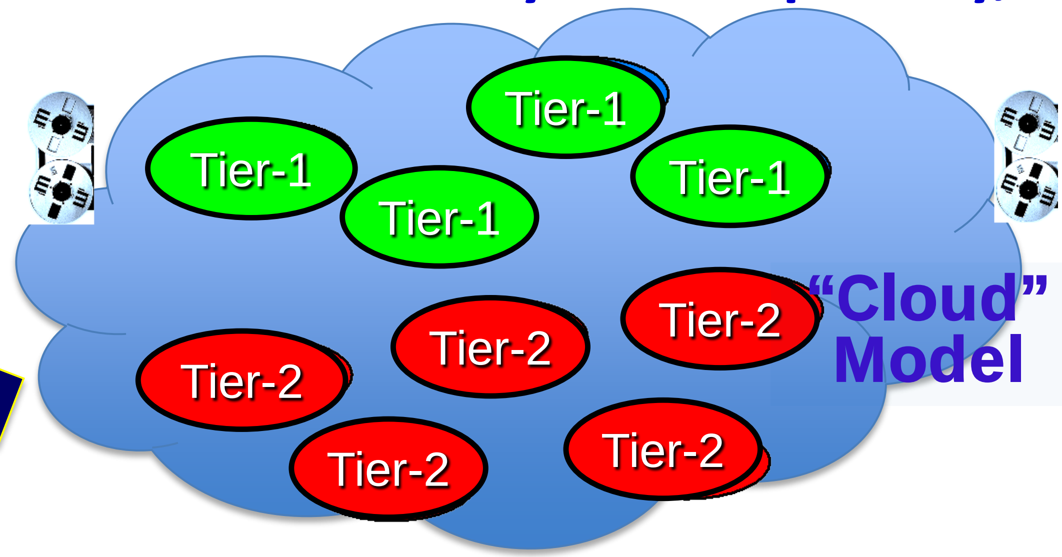
newman@hep.caltech.edu

Location Independent Access: Blurring the Boundaries Among Sites + Analysis vs Computing

- Once the archival functions are separated from the Tier-1 sites, the **functional difference between Tier-1 and Tier-2 sites becomes small [and the analysis/computing-ops boundary blurs]**
- **Connections and functions of sites are defined by their capability, including the network!!**

Maria Girone
CMS Computing





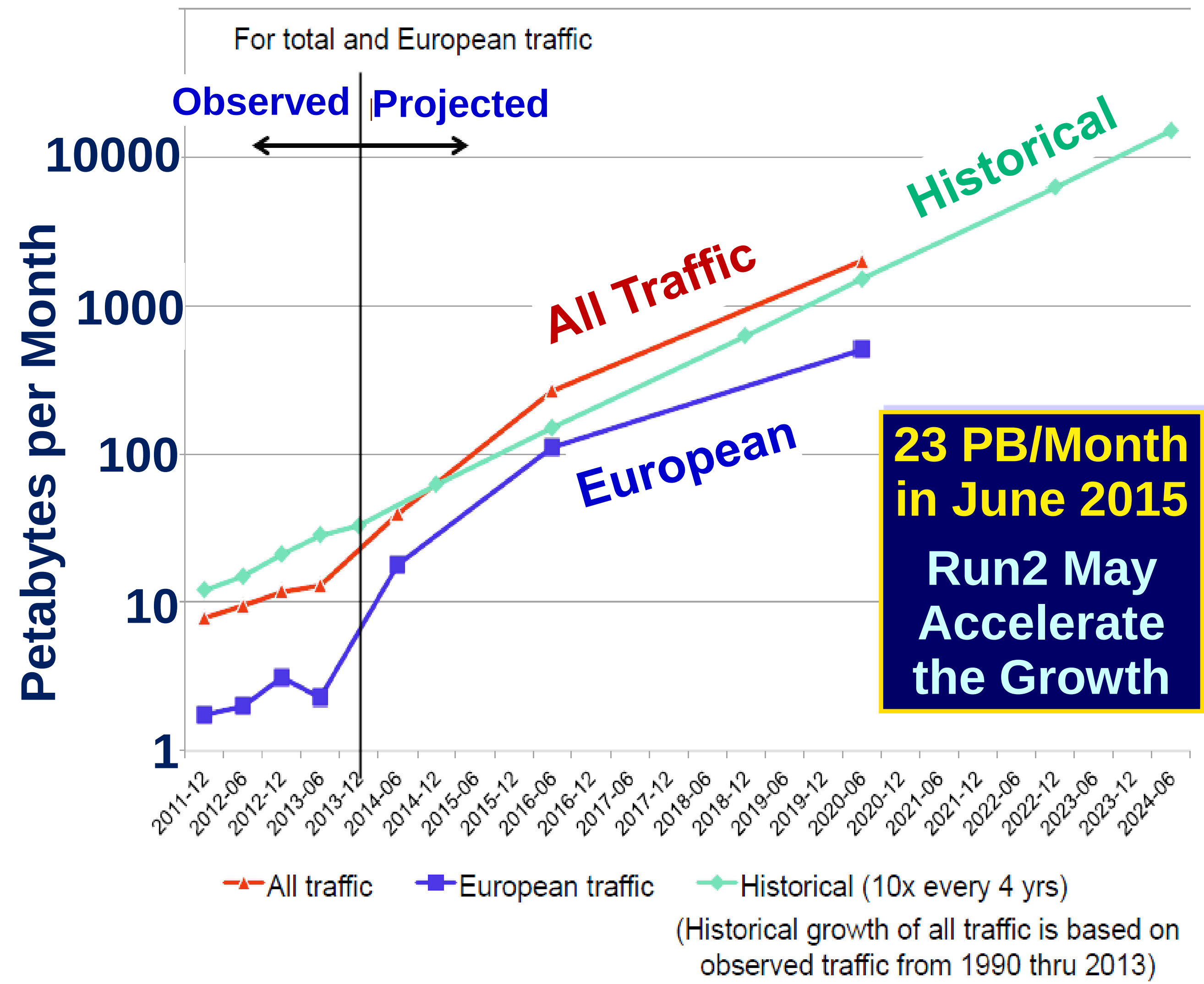
Run2: Scaling to 20% of data across the WAN: 200k jobs, 60k files, (100TB)/day

✷ + Elastic Cloud-like access from some Tier1/2/3 sites

ESnet Science projection to 2024
Compared to historical traffic

E. Dart
W. Johnston

Total traffic handled in Petabytes per Month



23 PB/Month in June 2015

Run2 May Accelerate the Growth

Projected Traffic Reaches 1 Exabyte Per Month. by ~2020

10 EB/Mo. by ~2024

Rate of increase follows or exceeds Historical trend of 10X per 4 Years

HEP traffic will compete with BES, BER and ASCR

Entering a new Era of Exploration and Discovery in Data Intensive Sciences

- We are entering a new era of exploration and discovery
  - In many data intensive fields, from HEP and astrophysics to climate science, genomics, seismology and biomedical research
- The largest data- and network-intensive programs from the LHC and HL LHC, to LSST and DESI, LCLS II, the Joint Genome Institute and other emerging areas of growth face unprecedented challenges
  - In global data distribution, processing, access and analysis
  - In the coordinated use of massive but still limited CPU, storage and network resources.
- High-performance networking is a key enabling technology for this research: global science collaborations depend on fast and reliable data transfers and access on regional, national and international scales

The Future of Big Data Circa 2025:

Astronomical or Genomical ? By the Numbers

PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195

Domains of Big Data in 2025. In each, the projected annual and storage needs are presented, across the data lifecycle

Basis: 0.1 to 2B Humans with Genomes, replicated 30Xs;

+ Representative Samples of 2.5M Other Species' Genomes

Data Phase	SKA	Twitter	YOU TUBE	GENOMICS	HL LHC
Acquisition	25 ZB/Yr	0.5–15 billion tweets/year	500–900 million hours/year	1 Zetta-bases/Yr	
Storage	1.5 EB/Yr	1–17 PB/year	1–2 EB/year	2-40 EB/Yr	2-10 EB/Yr
Analysis	In situ data Reduction	Topic and sentiment mining	Limited requirements		
	Real-time processing	Metadata analysis		Variant Calling 2×10^{12} CPU-h	
	Massive Volumes			All-pairs genome alignment 10^{16} CPU-h	0.065 to 0.2 X 10^{12} CPU Hrs
Distribution	DAQ 600 TB/s	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many at 10 MBps Fewer at 10 TB/sec	DAQ to 10 TB/s Offline ~0.1 TB/s

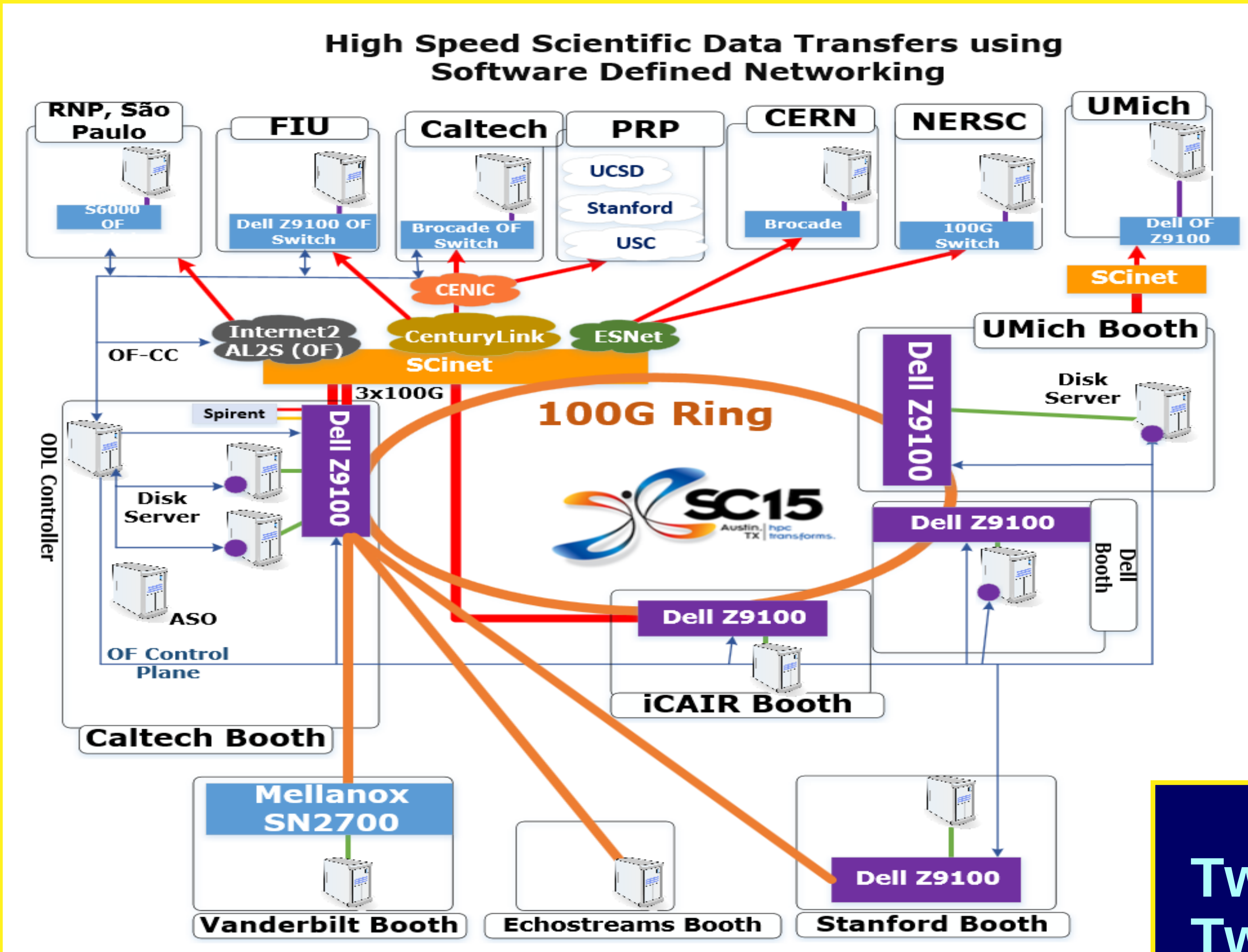
Conclusion: Genomics Needs Realtime Filtering/Compression Before a Meaningful Comparison Can Be Made

Servers at the Caltech Booth

Multi-100G DTNs, SDN, Machine Learning

- 2 400G DTNs: Dell Model R930 4U servers with E7 four socket CPUs; each with 4 100G NICs
- Third R930 server with 5 Model MX6300 Mangstor cards capable of 18+/12 R/W GBytes/sec
- Fourth R930 server with 4 Intel Model DC P3700 SSDs and 1 100G NIC
- 1 Supermicro server with 8 Intel Model DC P3700 SSDs, 2 40G Mellanox NICs (Connects to ESnet)
- 2 Supermicro 4U dual E5-2697 servers each with 3 100G NICs
- 3 SuperMicro (2U dual E5-2670) w/24 OCZ Vertex4 and Intel SSDs each
- 2 Echostreams server (4U and 2U, processors dual E5 2.2 GHz) each with 100G Mellanox NIC
- Echostreams/Orange Labs Server with 16 Tesla K80 GPUs: 100 Teraflops in 4U

SC15: SDN Driven Next Generation Terabit/sec Integrated Network for Exascale Science



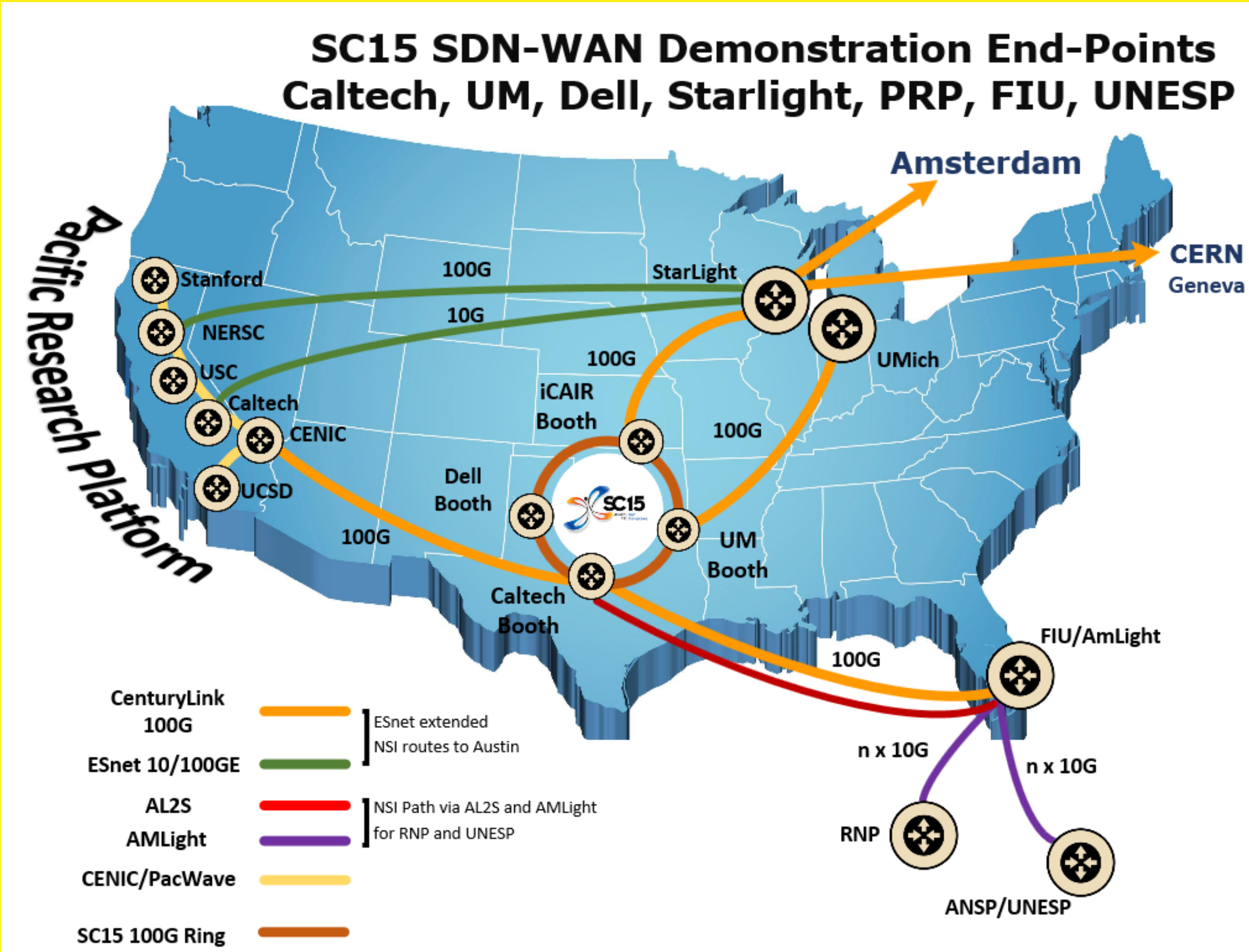
SDN-driven flow steering, load balancing, site orchestration Over Terabit/sec Global Networks

Consistent Operations with Agile Feedback: Supporting Major Science Flows Compatible with other Traffic

29 100G NICs
Two 4 X 100G DTNs
Two 3 X 100G DTNs
9 32 X100G Switches

Caltech HEP & Partners. Open Daylight Controller

SC15: SDN Driven Next Generation Terabit/sec Integrated Network for Exascale Science



Caltech HEP & Partners. Open Daylight Controller

SDN-driven flow steering, load balancing, site orchestration Over Terabit/sec Global Networks

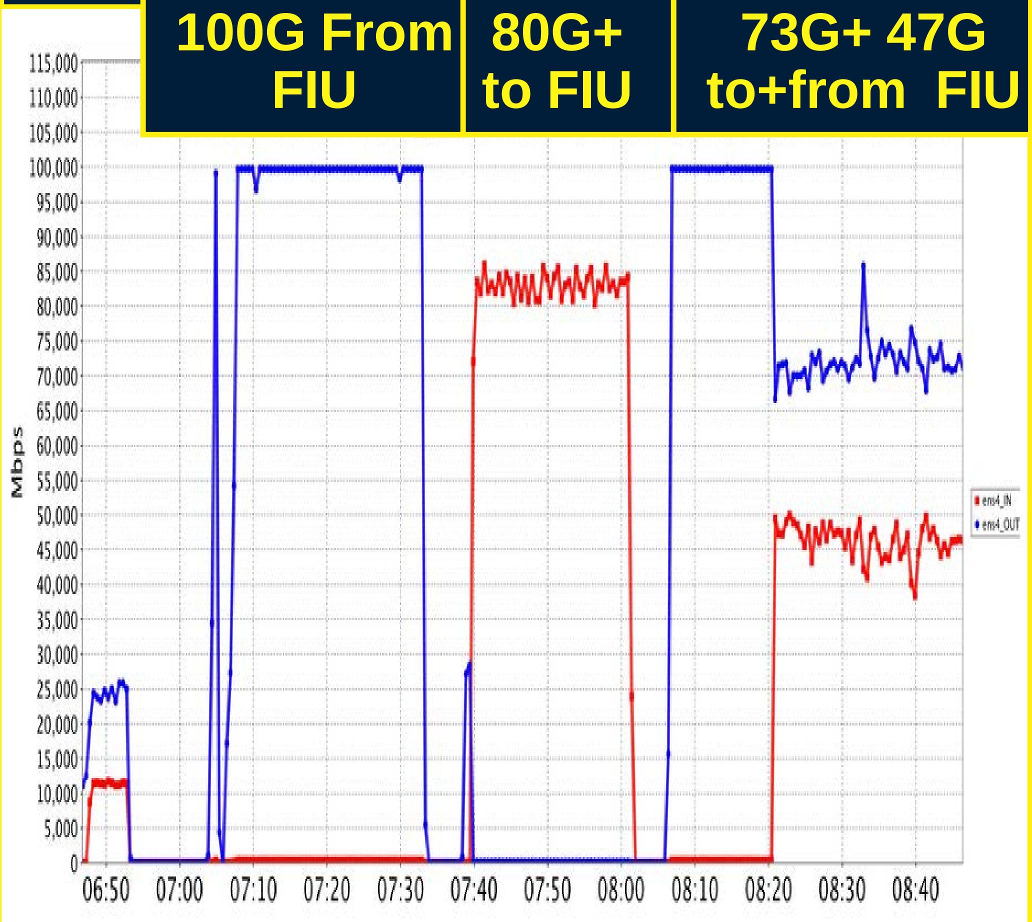
Consistent Operations with Agile Feedback: Major Science Flow Classes Up to High Water Marks

PetaByte Transfers to and From the Site Edges of Exascale Facilities With 400G DTNs

Mellanox and Qlogic 100G and Mellanox N X 100G NIC Results

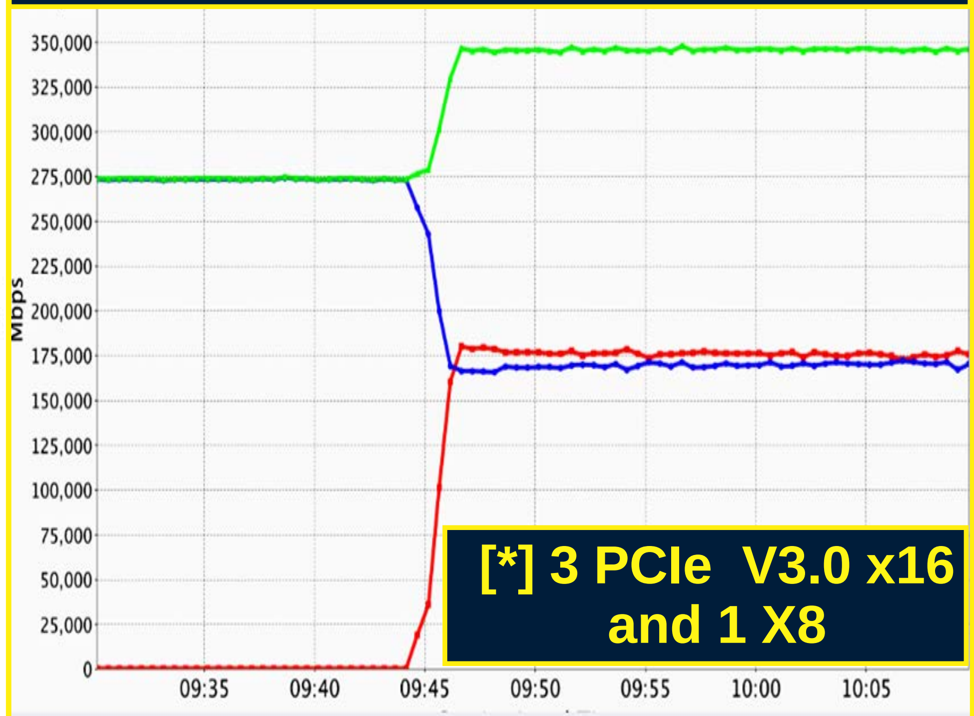
FIU – Caltech Booth – Dell Booth

100G From FIU | 80G+ to FIU | 73G+ 47G to+from FIU



4 X 100G Server Pair in the Caltech Booth

275G out; 350G in+out [*]
Stable Throughput

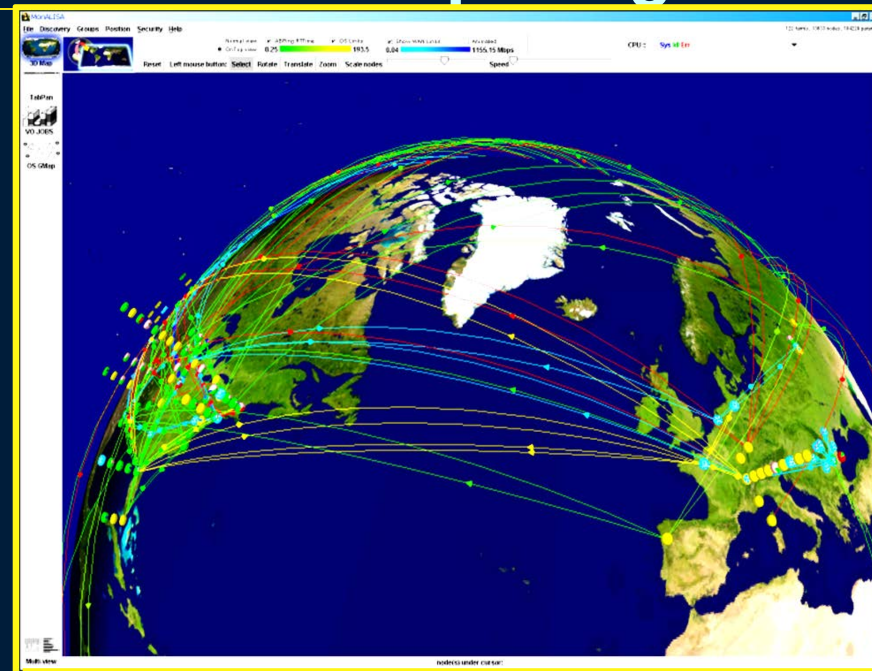


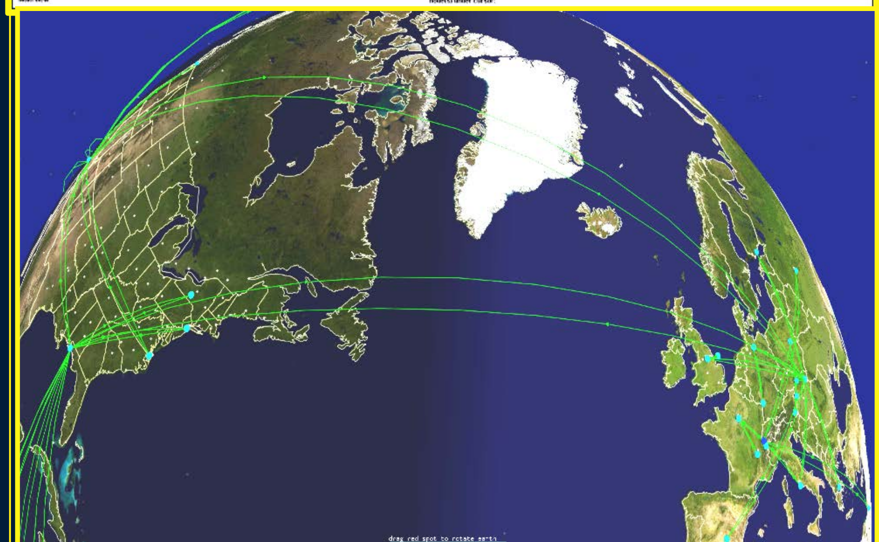
[*] 3 PCIe V3.0 x16 and 1 X8

Using Caltech's FDT TCP Application
http://monalisa.caltech.edu/FDT

Entering a New Era of Technical Challenges as we Move to Exascale Data and Computing

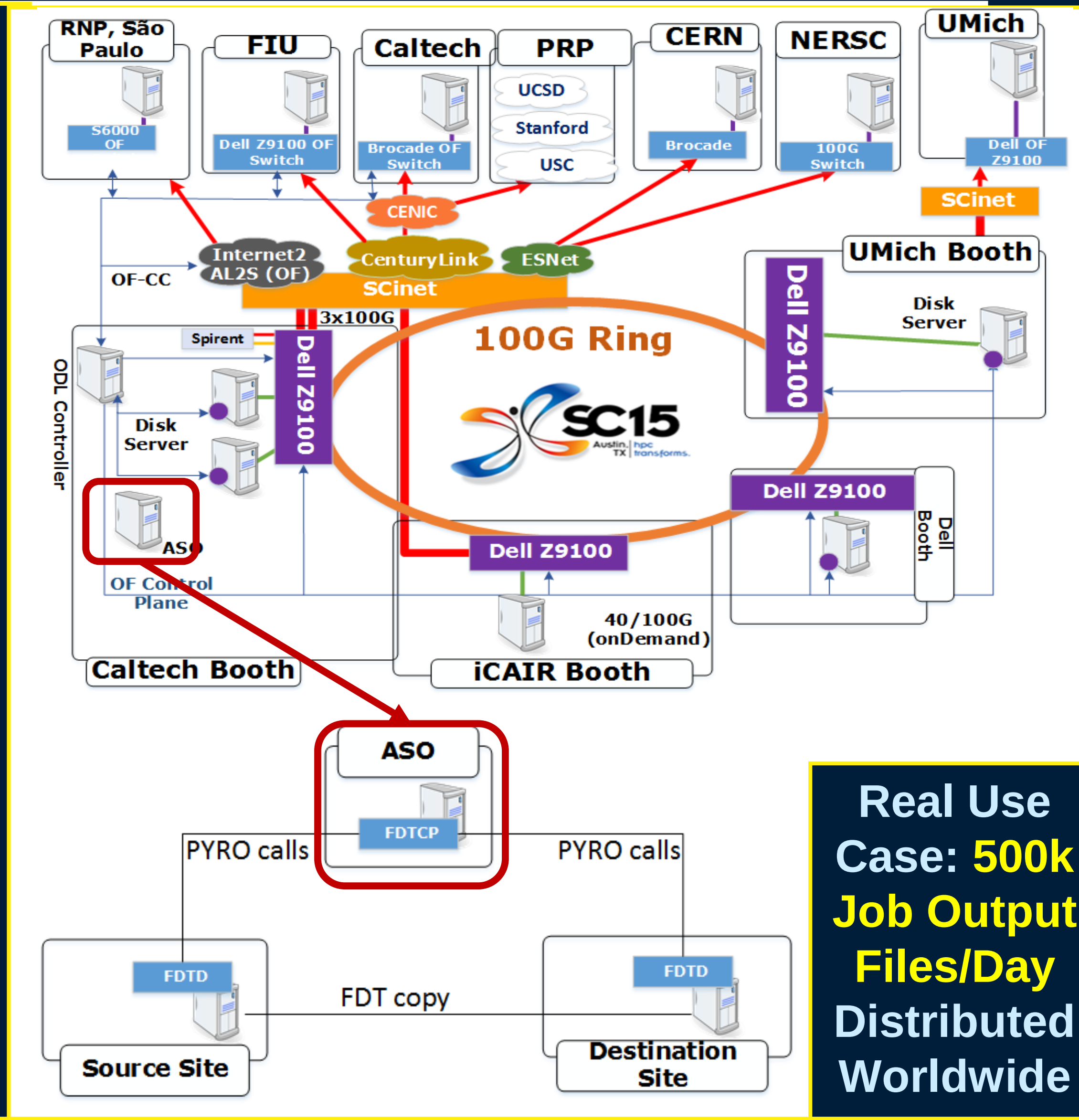
- **Beyond network capacity and reliability alone, the keys to future success are next generation systems able to:**
  - **Respond agilely to peak and shifting workloads**
  - **Accommodate a more diverse set of computing systems from the Grid to the Cloud to HPC**
  - **Coordinate the use of globally distributed computing and storage, and networks that interlink them**
    - **In a manner compatible across fields sharing common networks**
- **The complexity of the data, and hence the needs for CPU power, will grow disproportionately: by a factor of several hundred during the same period**





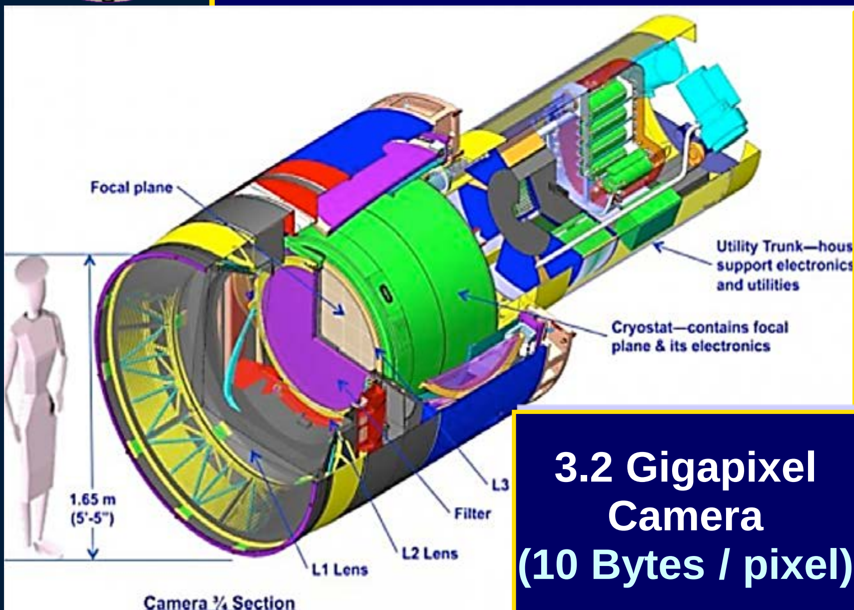
CMS at SC15: Asynchronous Stage Out 3rd Party Copy Demonstration

- All control logic in ASO:
  - Group multiple file transfers per link
  - Controls number of parallel transfers
  - Transparent for ASO integration
- Only FDT daemon has to be installed on storage site
- Tests between end-hosts at Caltech, Umich, Dell booths and outside: FIU, Caltech, CERN, Umich
- PetaByte transfers from multiple sites to multiple locations



Real Use Case: 500k Job Output Files/Day Distributed Worldwide

LSST + SKA Data Movement
Upcoming *Real-time* Challenges for Astronomy



3.2 Gigapixel Camera (10 Bytes / pixel)

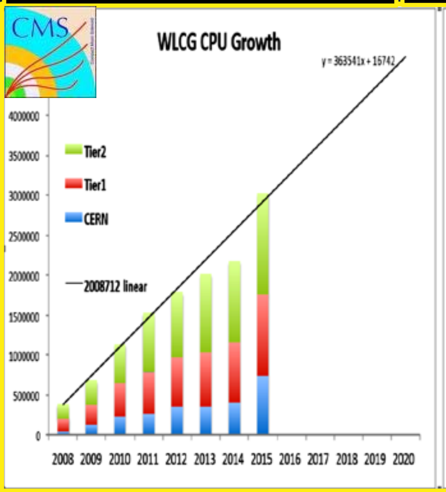
SKA

- **Planned Networks**: Dedicated 100G for image data, Second 100G for other traffic, and 40G for a diverse path
- **Lossless compressed Image size = 2.7GB** (~5 images transferred in parallel over a 100 Gbps link)
  - **Custom transfer protocols for images (UDP Based)**
- **Real-time Challenge: delivery in seconds to catch cosmic “events”**
- **+ SKA in Future: 3000 Antennae covering > 1 Million km2;** *15,000 Terabits/sec to the correlators ➡ 1.5 Exabytes/yr Stored*

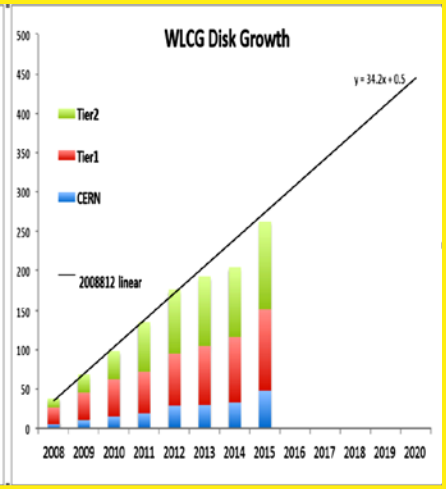
CMS Offline Computing Requirements

HL LHC versus Run2 and Run1 [*]

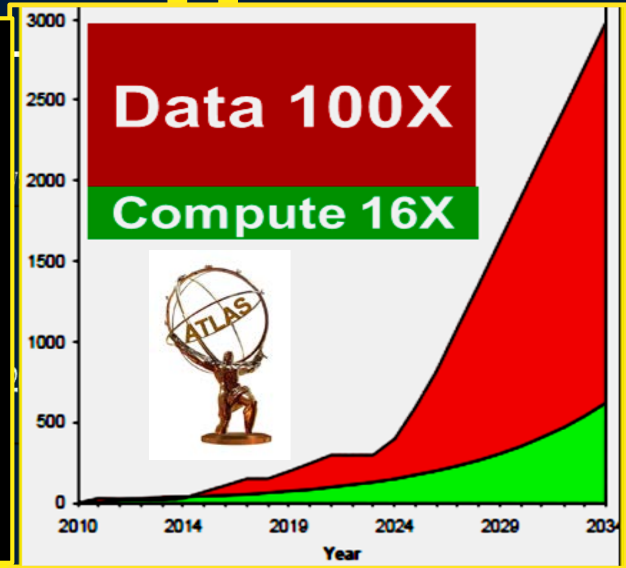
+ ~36k cores/Yr



+ ~34 PBytes/Yr



- Ratios in Computing and Storage for Run 2/Run1 are ≈ 2X.
- Hence HL-LHC to Run1 CPU: 130X to 400X



CPU Requirements Projections

- Projected CPU Needs:
  HL LHC/Run2 = 65 to 200X
- Anticipated increase in CPU resources at fixed cost/year: 8X
- Anticipated code efficiency improvements: 2X
- Projected shortfall at HL LHC
  4X to 12X

Storage Requirements Projections

- Projected Events:
  HL LHC / Run2 = 5 to 7.5X
- Event Size:
  HL LHC / Run2 = 4 to 6X
- Anticipated growth in Storage
  HL-LHC / Run2: 20-45X
- Projected shortfall at HL LHC
  3X or More

[*] CMS Phase2 Technical Proposal: https://cds.cern.ch/record/202088

HEP Collider HPC Use, Prospects and Wishes

Tom Lecompte (Argonne) at the Exascale Workshop

❖ Computing to reach the Science Goals: Argonne LCF Use

To MIRA

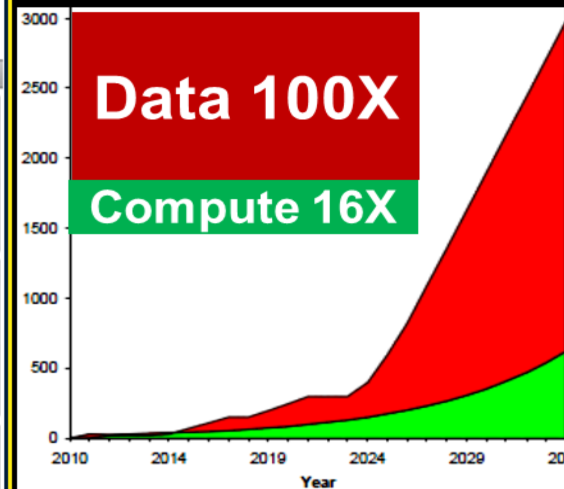
- Mira: 65M Hrs
  Compare the
  Grid: 1B Hours
- 2 FTEs
- Equal to the 7th largest "country" in CPU power in ATLAS in 2015
- Focus: Generators

- Simulation next
- Enabling "extra dimensions" in HEP Analysis

ALPGEN on MIRA

- 256k/768k Cores
- Code Improved 23X: 1 core went from 1/15 to 1.5X a Grid core
- 6-8X the ATLAS Grid CPU when running

Issues for the HL LHC



Meeting the CPU and data handlng needs

Adapting HEP codes + SW frameworks

Beyond application software alone: A New Class of System

An excellent very promising start. A lot of work remains

Key Developments on the HEP Side

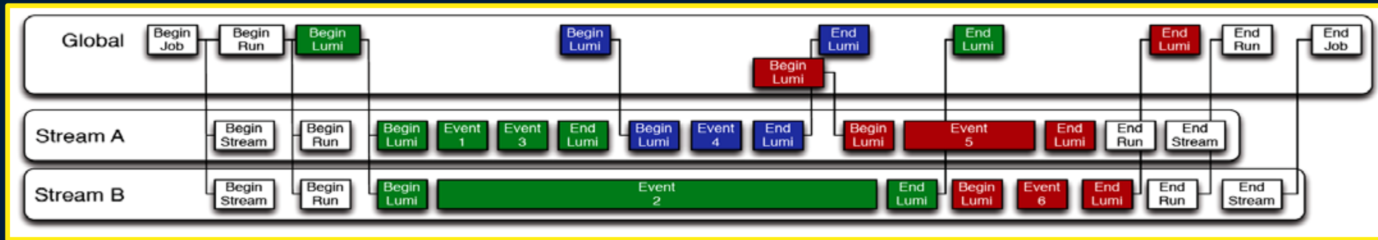
Enabling the Vision: Coherent Parallel Architectures

- **We need to recast HEP's code and frameworks for the highly parallel, energy efficient architectures (GPU, Knights Landing, etc.) of modern HPC systems**

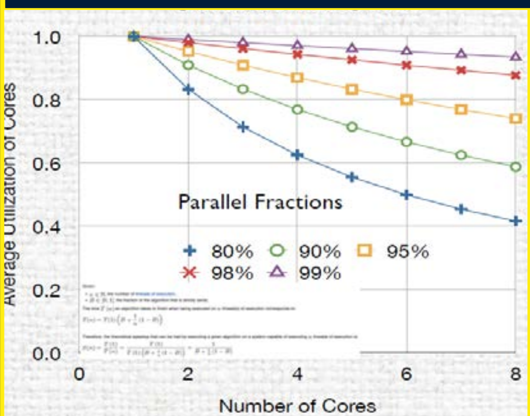
***Significant progress* in specific HEP areas exists**

- **CMS threaded memory-efficient concurrent framework for multicore CPUs**

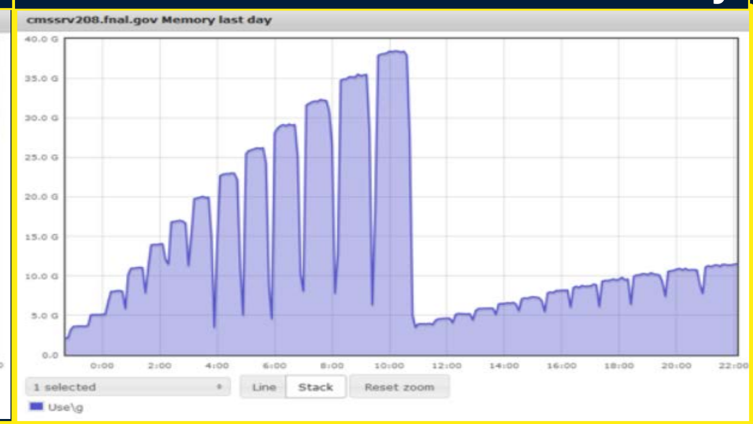
CMS Multithreaded Reconstruction Framework
E. Sexton-Kennedy at CHEP2014



90% efficient with 16 threads ➡ reco code is 99.3% parallel



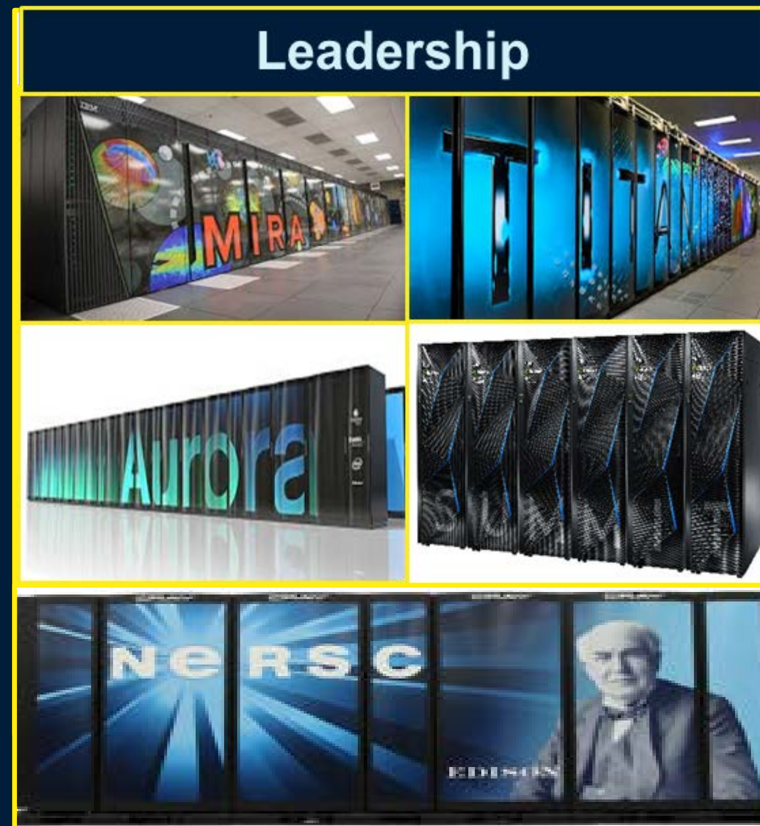
New code saves 2/3 of the memory



- **ATLAS generators have successfully run on (all of) MIRA (100M events in 1M threads); Looking towards Aurora [Tom LeCompte ➡]**

Exascale "CSN" Ecosystems for Next-Generation Data Intensive Science

- The opportunity for HEP (CMS example):
  - CPU needs will grow 65 to 200X by HL LHC
  - Dedicated CPU that can be afforded will be an order of magnitude less; even after code improvements on the present trajectory
- DOE ASCR/HEP Exascale Workshop:
  - Identified key opportunities for harnessing the special capabilities of ECFs
  - Exposed the favorable outlook and issues for HEP to take this key step + meet the needs
  - Highlighted the *Network Dimension*
- Important added benefits to HEP + ASCR, the facilities, programs and the nation
  - Shaping the future architecture and operational modes of ECFs
  - Folding LCFs into a global ecosystem for data intensive science
  - Developing a "modern coding workforce"
  - Enabling many fields to "think out of the box"

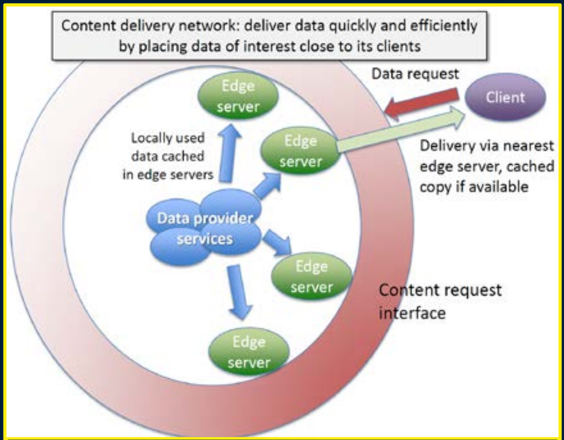


A favorable HEP platform:

- LHC experiments are gearing their S&C operations for more flexible use of diverse resources: Grid, Cloud, HPC

LCF-Edge Data Intensive Systems (LEDIS) Operational Model

- ❒ In the context of a new HEP – LCF – ESnet partnership for **Joint system and architecture development**
- ❒ Data brought to LCF edge ~petabyte chunks: **Delivery in ~2 hrs at 1 Tbps**
  - ❒ **Far enough in advance:** chunks ready and waiting in a buffer pool
- ❒ Using secure systems at the site perimeter: **Security Efforts (human and AI)** can be focused on a limited number of entities **(proxies)**
  - ❒ **Keeping manpower + risk at acceptable levels**
- ❒ Multiple chunks for different stages of the workflow
  - ❒ **Each chunk's provenance + attributes identified**
  - ❒ **Examples:** Input/Output Data size, memory, CPU to IO ratio; delivery deadline, authorization level
- ❒ Enables matching to appropriate HPC subsystems, **to meet the needs while operating at high efficiency**
- ✸ Conceptual Extension: **Caching in the Network, or at nearby HEP Lab Sites; as in** "Data Intensive" CDNs (or NDNs)
  - ❒ Adapting to the future Internet architecture **that may emerge**



Follow on to a CDN Model
T. Wenaus BNL/ATLAS

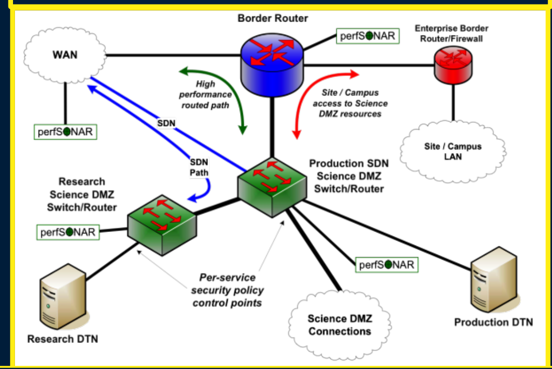
Key Developments from HPC Facility Side

Enabling the Vision: ECF Architecture

- ❒ Developing appropriate system architectures in hardware + software that meet the needs
  - ✸ Edge clusters with petabyte caches
    - ✸ Input + output pools: ~10 to 100 Pbytes
  - ✸ A handful of proxies at the edge
    - ✸ To manage and focus security efforts
  - ✸ Identifying + matching HEP units of work to specific sub-facilities adapted to the task
  - ✸ Extending Science DMZ concepts
    - ✸ Enabling 100G to Tbps SDNs with Edge/WAN Coordination + DTN Autoconfiguration
  - ✸ Site-Network End-to-End Orchestration
    - ✸ Efficient, smooth petabyte flows

Leadership

Next Gen Science DMZ



➡ Dynamic agile systems that learn to adapt to peaking workloads

Convergence and Collaboration
Tackling the Larger Mission

- Empowering Data Intensive Science across multiple fields through efficient, manageable use of national & global infrastructures up to high occupancy levels, including multi-pathing
- Using SDN-driven coordinated use of computing, storage and Network resources for efficient workflow
- Enabled by Pervasive End-to-end Monitoring
- Consistent Operations: Networks ⇔ Science Programs; with feedback
- Key Concepts and Technologies for Success:
  - Dynamic circuits for priority tasks, with Transfer Queuing, Deadline scheduling, Efficient worldwide distribution and sharing
  - Classes of Service by flow characteristics, residency time
  - Load balancing, hotspot resolution, strategic redirection
  - State-based error propagation, localization, resolution
  - SDN driven Intent-based deep site-network orchestration functions
- System Level Optimization Using Machine Learning

SDN in SDN-NGenIA and SENSE

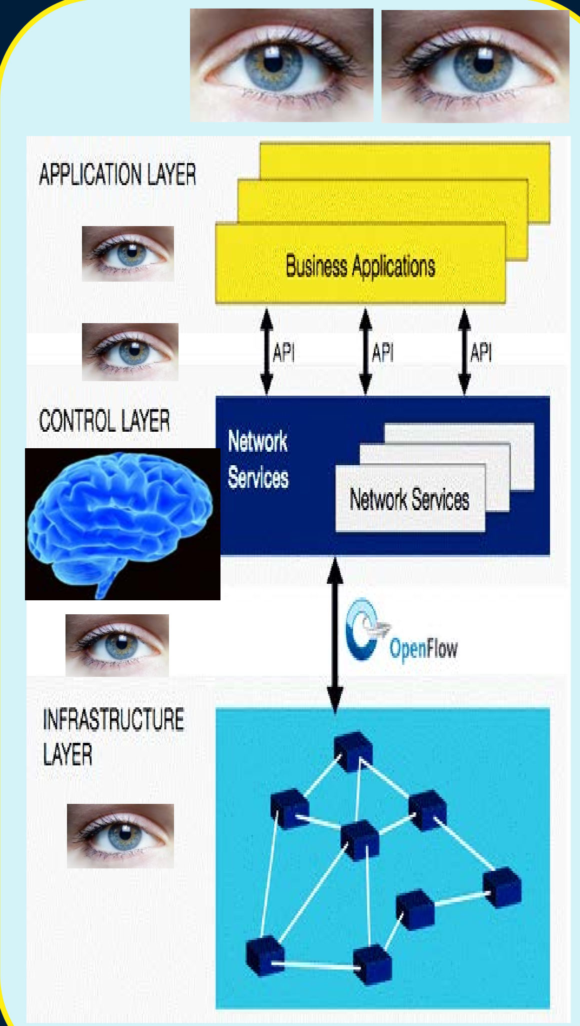
Ideas Building on Caltech/Esnet/FNAL Experience

Vision: Distributed computing environments where resources can be deployed easily and flexibly to meet the demands of data-intensive science, **giving transparent access to an integrated system of enormous computing power**

SDN is a natural pathway to this vision: separating the functions that control the flow of network traffic, **from the switching infrastructure that forwards the traffic itself** through open deeply programmable "controllers".

With many benefits:

- Replacing stovepiped vendor HW/SW solutions by open platform-independent software services
- Imagining new methods and architectures
- Virtualizing services and networks: lowering cost and energy, with greater simplicity

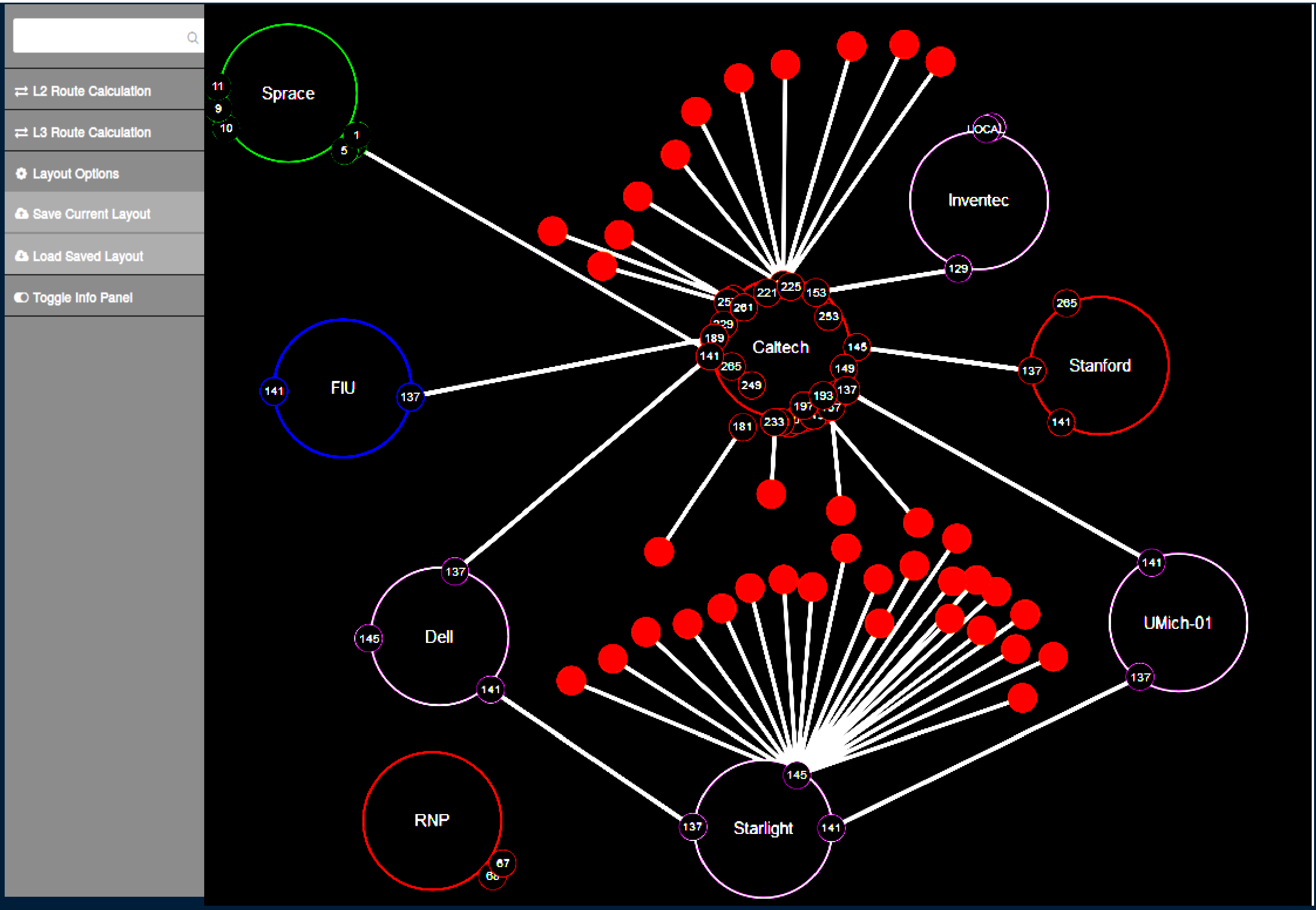


opennetworking.org

A system with built in intelligence

Requires excellent monitoring at all levels

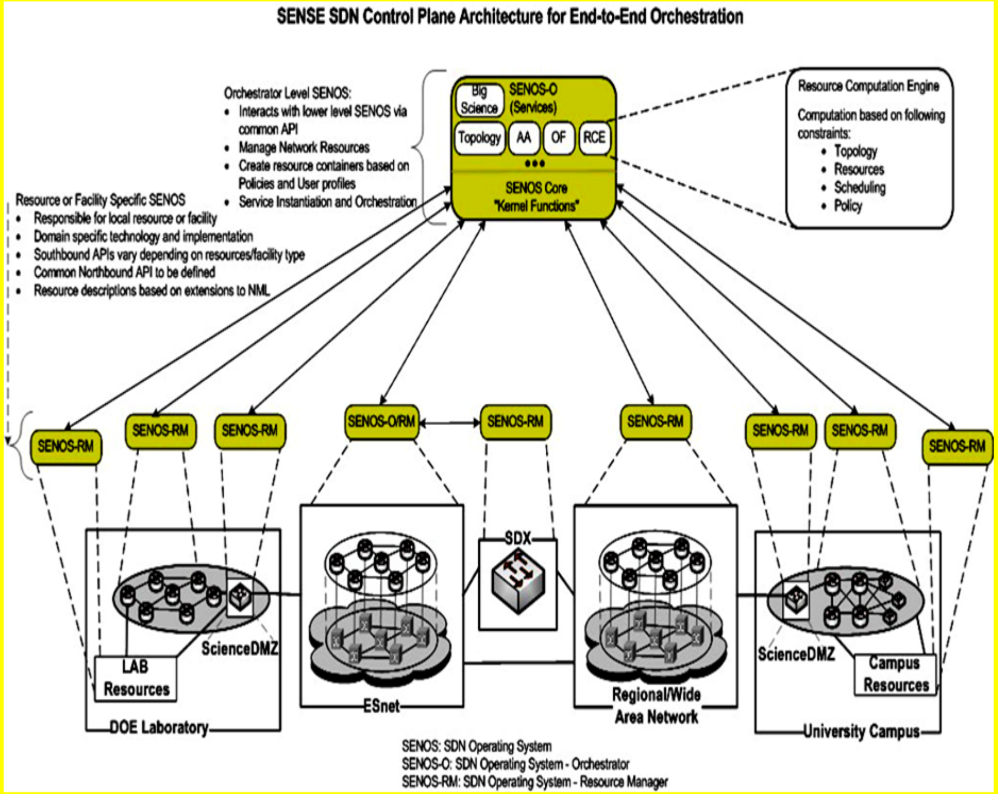
SC15: SDN Driven Terabit/sec
Live OF Network Topology for Directing Flows



SENSE: SDN for End-to-end Networked Science at the Exascale

ESnet Caltech Fermilab Argonne Maryland

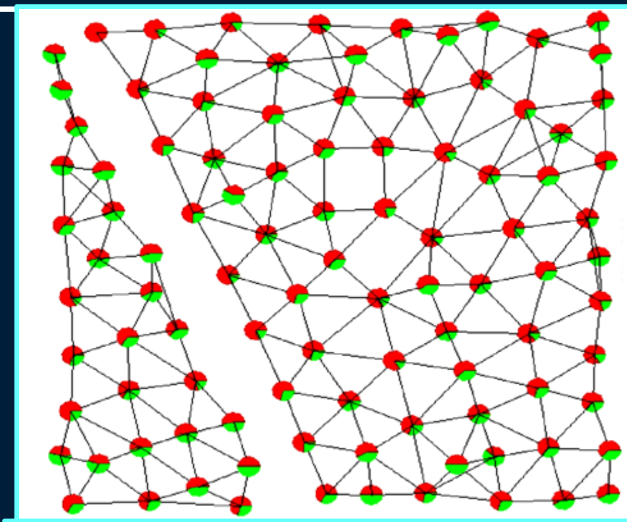
- **Mission Goals:**
  - Significantly improve end-to-end performance **of science workflows**
  - Enabling new paradigms: **creating dynamic distributed** 'Superfacilities'.

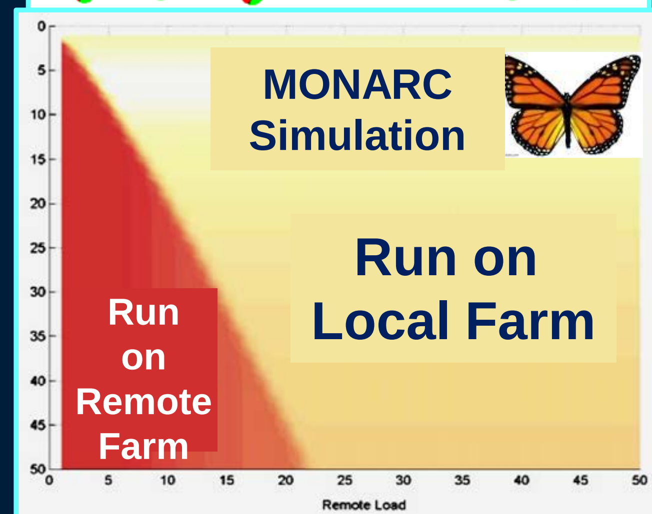


- **Comprehensive Approach:**
  An end-to-end SDN Operating System (SENOS), **with:**
  - Intent-based interfaces, **providing intuitive access to intelligent SDN services**
  - Policy-guided E2E orchestration **of resources**
  - Auto-provisioning **of network devices and Data Transfer Nodes**
  - Network measurement, analytics and feedback to **build resilience**

Key Developments from the HEP Side
Enabling the Vision: Machine Learning

- **Applying** Deep Learning + Self-Organizing systems methods **to optimize LHC workflow**
  - Unsupervised: **extract key variables/functions**
  - Supervised: **to derive optima**
  - Iterative and model based: **to find** effective metrics **and** stable solutions [*]
- **Complemented by** game theory methods, modeling and simulation
- Shown to be effective to solve traffic, communications and workflow problems
- **Starting with** logged monitoring information
- **Progressing to** real-time agent-based pervasive monitoring



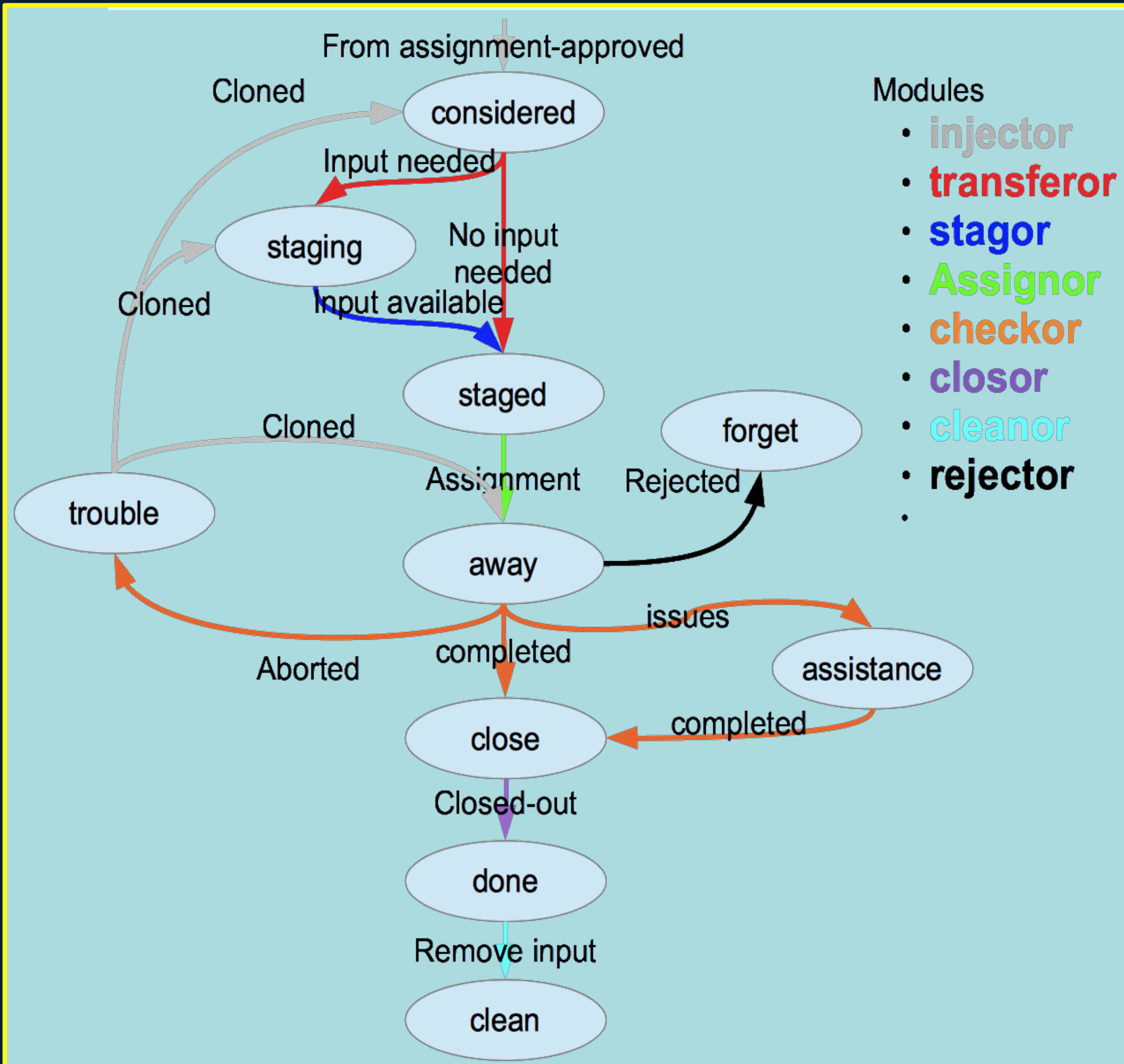


Self-organizing neural network for job scheduling in distributed systems

[*] T. Roughgarden (2005). *Selfish routing and the price of anarchy*

Computing Operation Automation

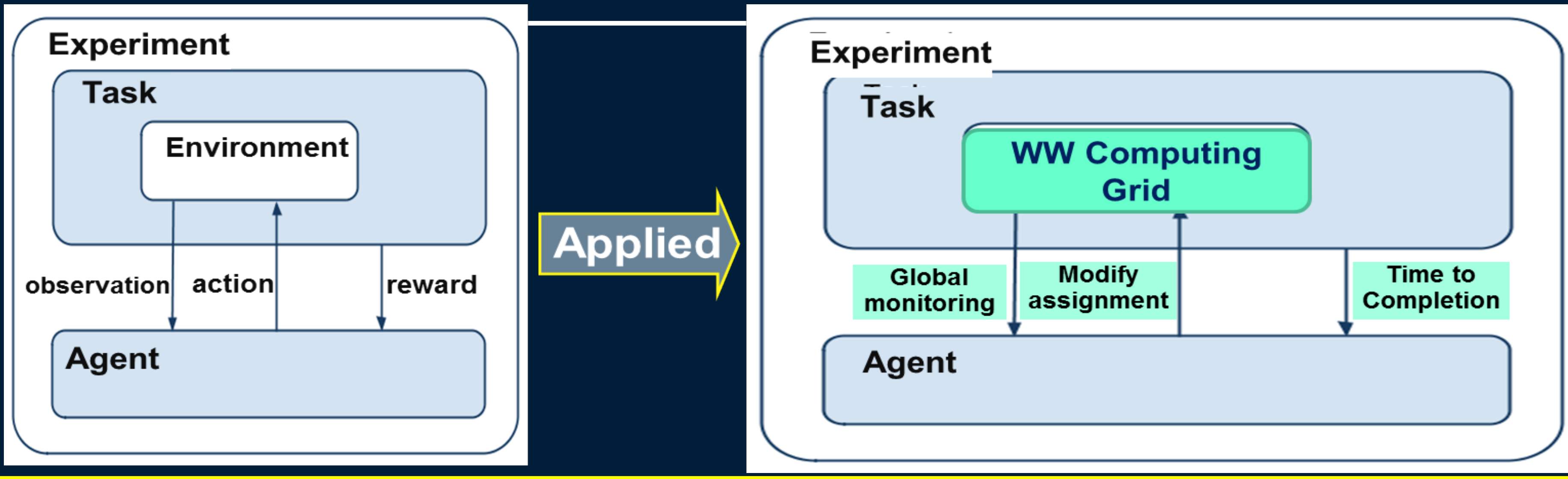
Example of a Model (State Machine)



- Fully automate handling of production requests
  - Pre-defined simple rules of placement
- Automation of sanity check and final delivery
  - Amount of operator work reduced
- Now possible to handle larger, more diverse resources smoothly

Computing Optimization R&D

Machine Learning Coupled to Modeling and Simulation



- Learn complex models using deep learning with monitoring data and the chosen metric(s)
- Use simulations together with game theory techniques or a reinforcement learning method to find optima
  - Variations: evolve towards the metrics yielding stable solutions with good throughput
  - Balancing among max throughput, balanced resource use, predicability of time to completion (predictable workflow) etc.
- Steering computing, storage and network elements like robot arms

Networks for HEP and Global Science

Our Journey to Discovery

- Run 1 brought us a centennial discovery: the Higgs Boson
- Run 2 will bring us (at least) greater knowledge, and perhaps greater discoveries: *Physics beyond the Standard Model.*
- *Advanced networks will continue to be a key* to the discoveries in HEP and other data intensive fields of science and engineering
- Technology evolution might fulfill the short term needs
- *Near Term Challenges:* A new net paradigm including the global use of circuits will need to emerge during LHC Run2 (in 2015-18)
- *New approaches + a new class of global networked systems to handle Exabyte-scale data are needed [LHCONE, DYNES, ANSE, OliMPS; SENSE+SDNNGenIA]*
- *Worldwide deployment of such systems in ~2020-24 will be:*
  - Essential for the High Luminosity LHC HL-LHC
  - A game-changer, with global impact, shaping both research and daily life

Data Intensive Exascale Facilities for Science
Deep Implications

Adapting Exascale Computing Facilities to meet the highest priority needs of data intensive science, including high energy physics as a first use case (to be followed by others)
will have profound implications:

- ❒ Empowering the HEP community to make the anticipated next and future rounds of discoveries
- ❒ Encouraging, and provoking the US scientific community to Think "top down" (Out of the Box) as well as "bottom up"
  - ❒ Envisioning a new scale; new applications, methods; and a new overall approach to science
  - ✸ Especially: in the face of an emerging discovery and the exploration of its aftermath
- ✸ HEP is a natural partner and thought co-leader in this process, and in the achievement of this goal

Data Intensive Exascale Facilities for Science
Deeper Implications

- Bringing these facilities into the ecosystem of globally distributed information and knowledge sources and sinks
  - The hallmark of science, research and everyday life this century
- Will open new avenues of thought and new modes of the pursuit of knowledge in the most data intensive fields
  - By responding to petascale inquiries on human time scales, irrespective of location
  - Bringing our major networks, once again, into sharp focus
- This will broaden the function and architecture of ECFs and ultimately shape them in future generations
  - While also shaping the leading edge of "modern computing and networking"
- And place the US science community in a new position of leadership
  Being the first to cross this conceptual threshold

THANK YOU!

Harvey Newman
newman@hep.caltech.edu