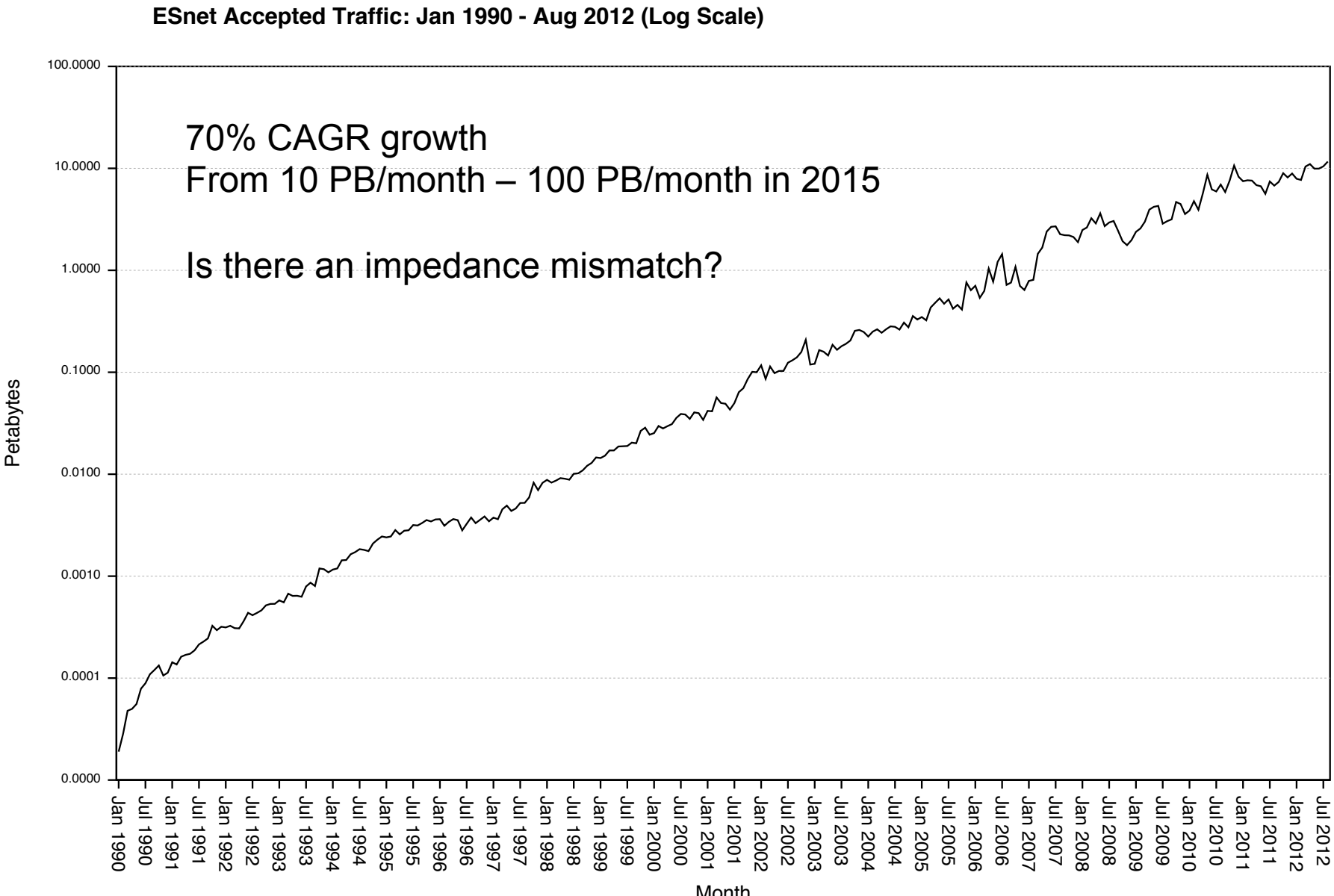
Programmable Information Highway (with no Traffic Jams)

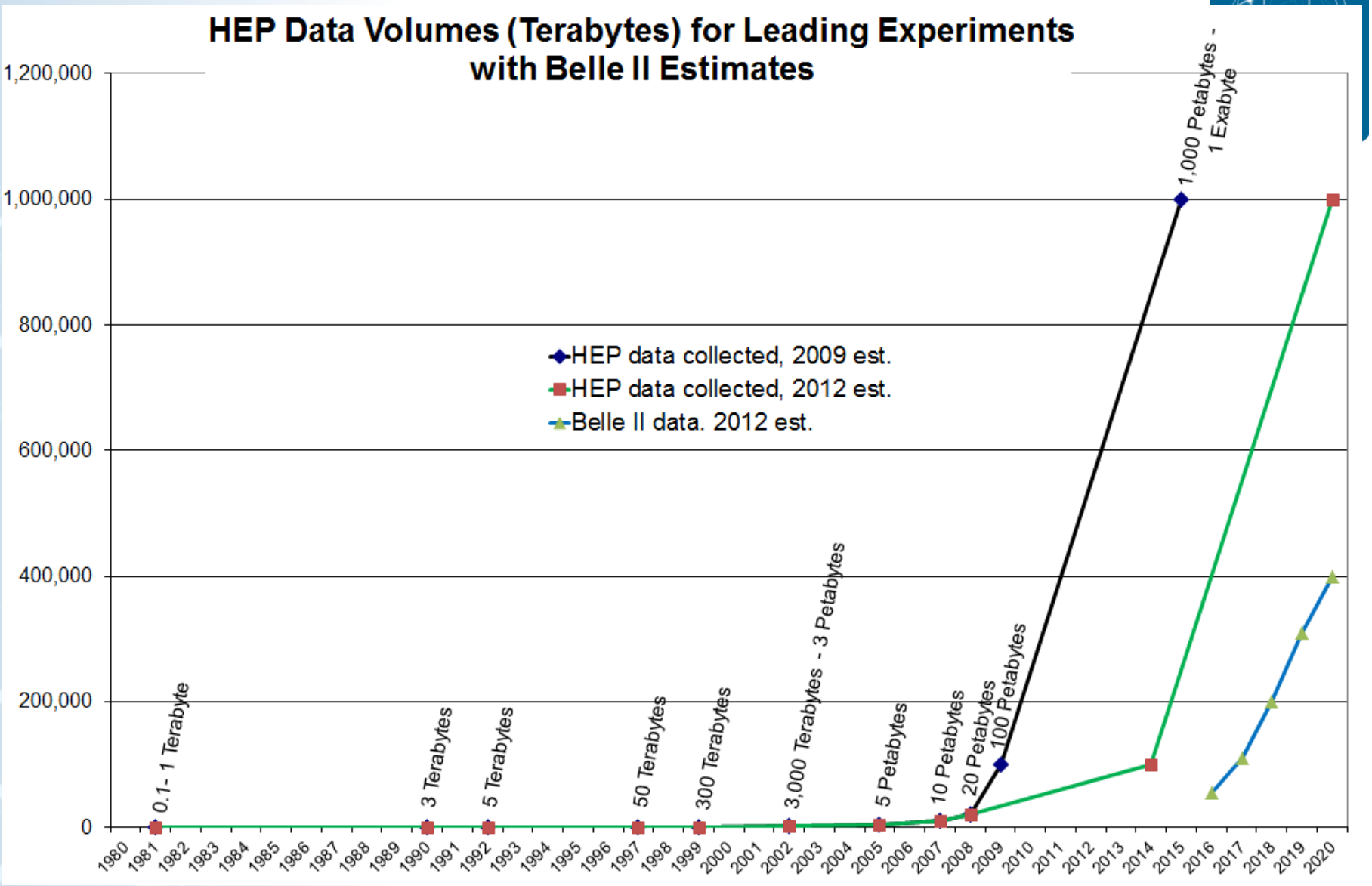
Inder Monga

Energy Sciences Network
Scientific Networking Division
Lawrence Berkeley National Lab

Exponential Growth

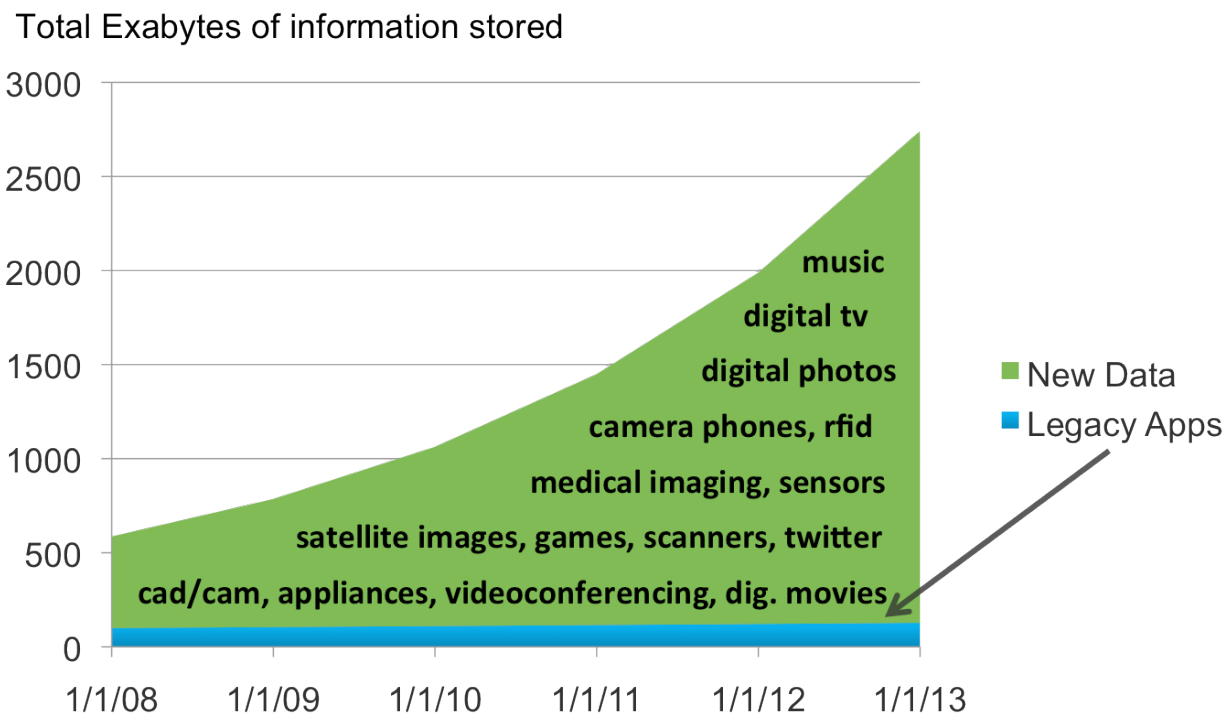


HEP as a Prototype for Data-Intensive Science

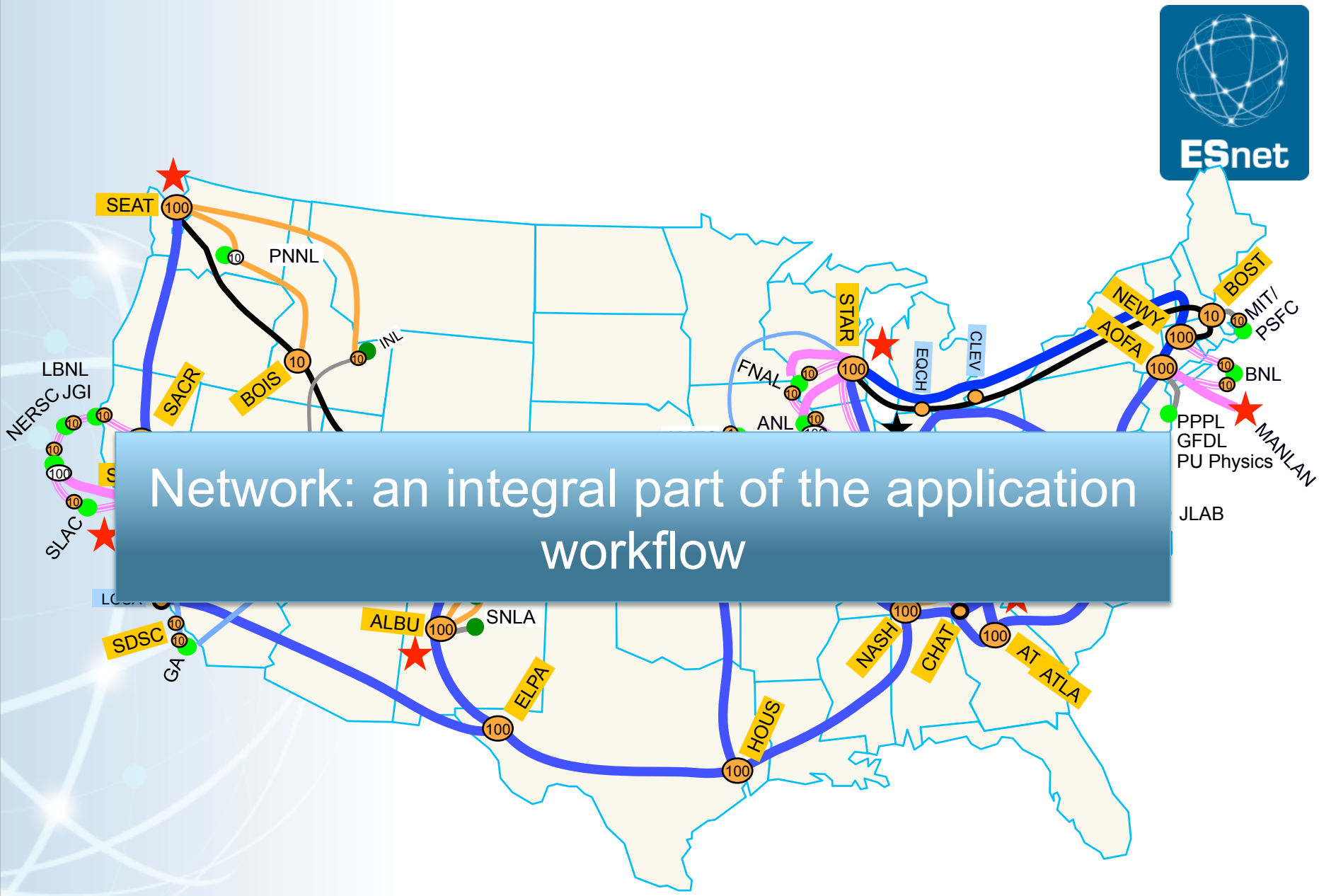


Data courtesy of Harvey Newman, Caltech, and Richard Mount, SLAC and Belle II CHEP 2012 presentation

Moving data is going to be a way of life

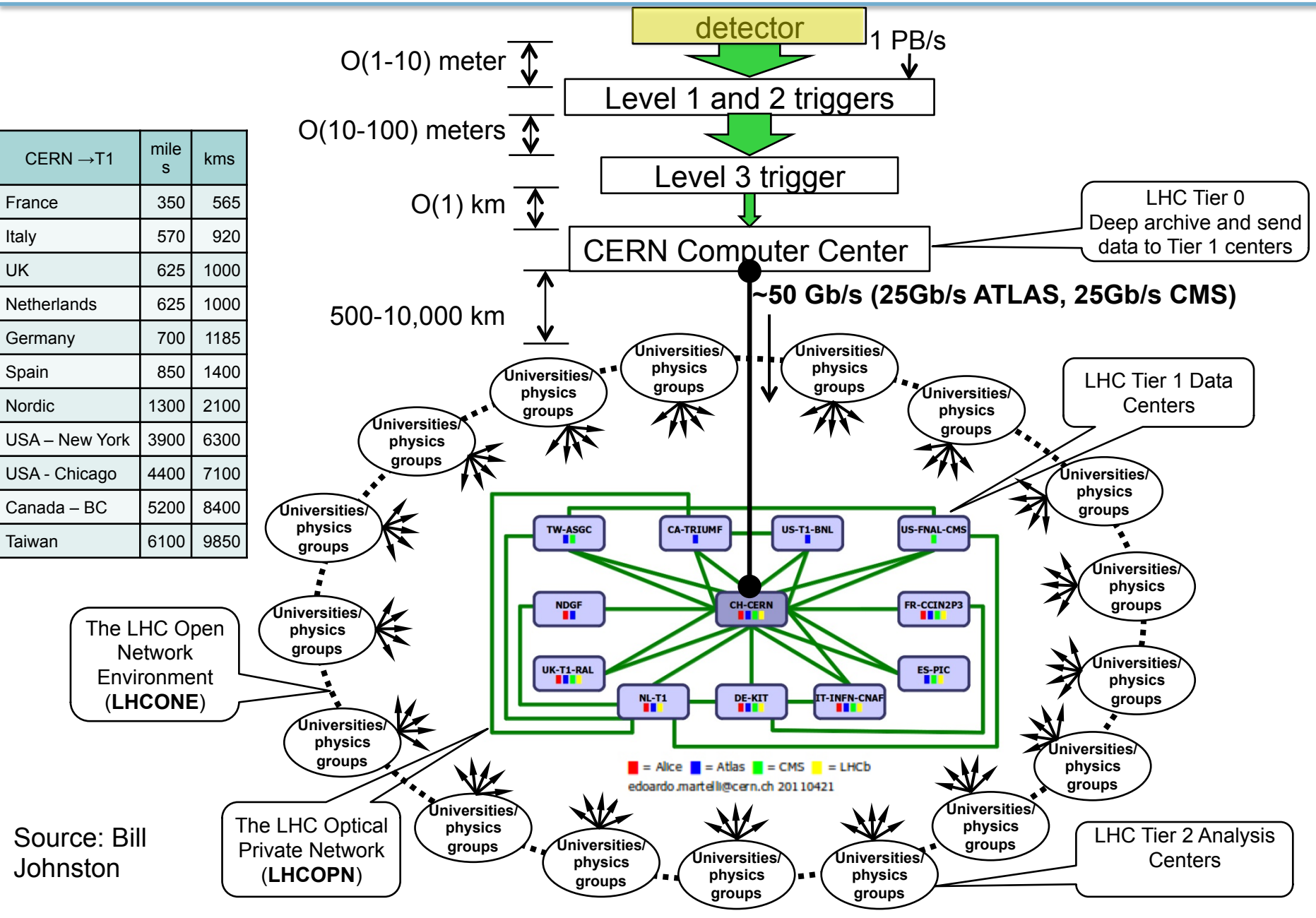


Source: The Information Explosion , 2009


ESnet
SEAT
PNNL
INL
LBNL
NERSC
JGI
SACR
BOIS
SLAC
FNAL
ANL
STAR
EQCH
CLEV
NEWY
AOFA
BOST
MIT/
PSFC
BNL
PPPL
GFDL
PU Physics
MANLAN
JLAB
SDSC
GA
ALBU
SNLA
ELPA
HOUS
NASH
CHAT
ATLA
Network: an integral part of the application workflow

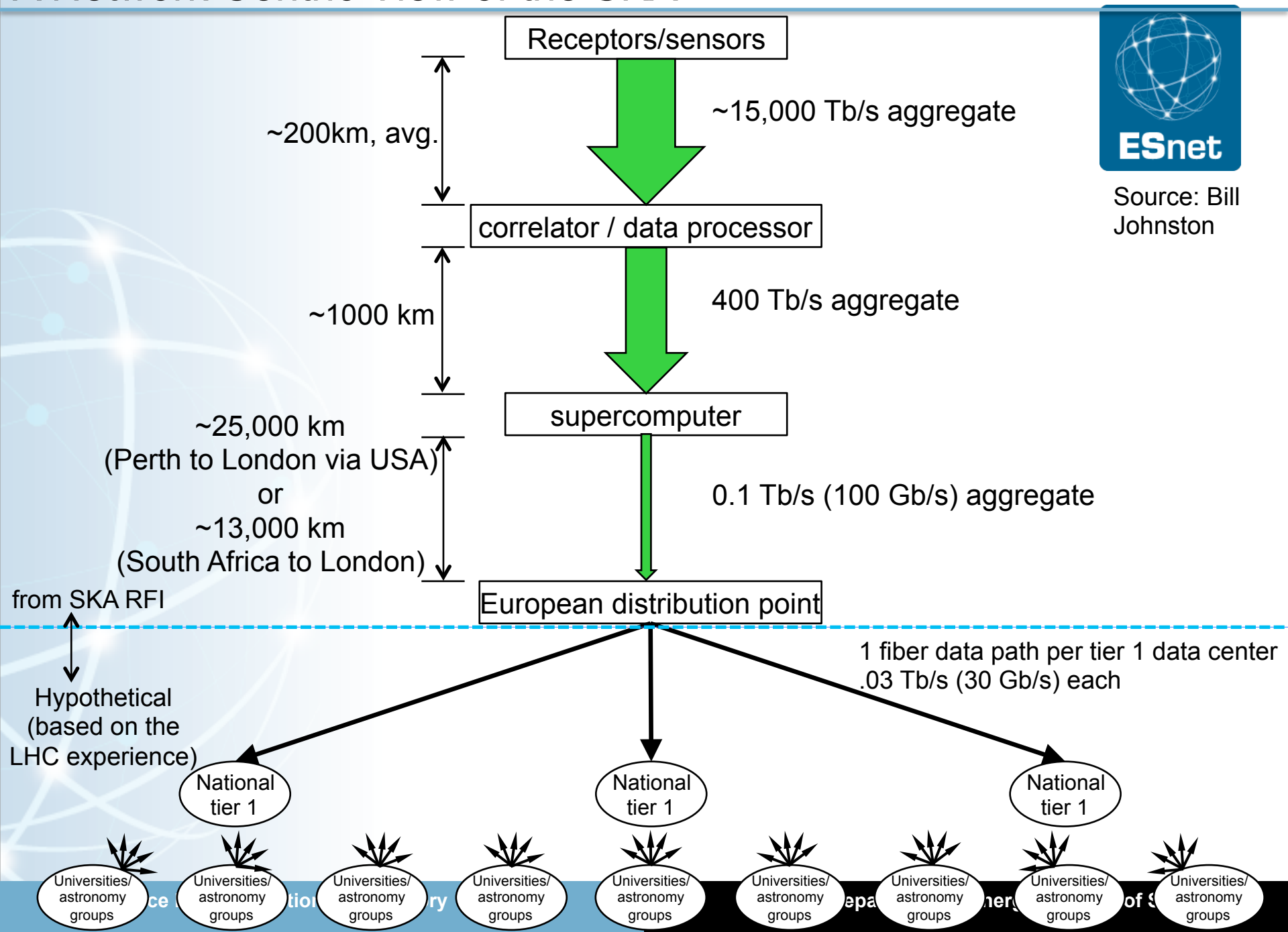
A Network-Centric View of LHC

CERN →T1	miles	kms
France	350	565
Italy	570	920
UK	625	1000
Netherlands	625	1000
Germany	700	1185
Spain	850	1400
Nordic	1300	2100
USA – New York	3900	6300
USA - Chicago	4400	7100
Canada – BC	5200	8400
Taiwan	6100	9850



Source: Bill Johnston

A Network Centric View of the SKA



ESnet

Source: Bill Johnston

New thinking changes the language of interaction

Infrastructure	Instrument
• Provides best-effort IP dialtone	• Adapts to the requirements of the experiment, science, end-to-end flow
• Average end-to-end performance, packet loss is fine	• Highly calibrated, zero packet loss end-to-end
• How much bandwidth do you need? (1G/10G/100G)	• What's your sustained end-to-end performance in bits/sec? Can I get the same performance anytime?
• Ping works, you are all set, go away	• Tuned to meet the application's workflow needs, across network domains

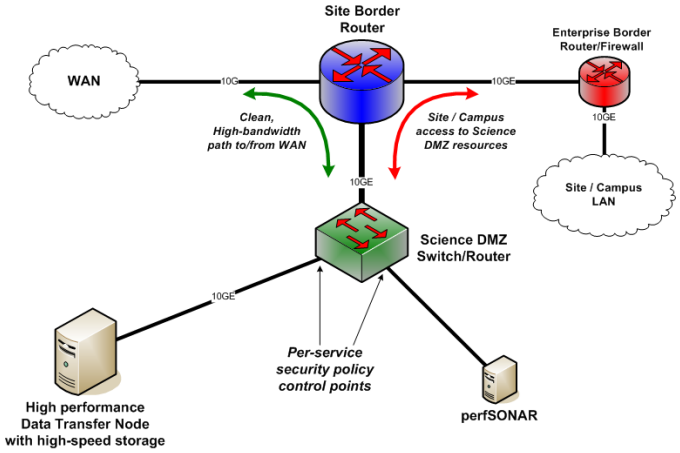
Adjectives for Network as a Instrument (Naal)

- End-to-End
- Programmable
- Simple
- Predictable

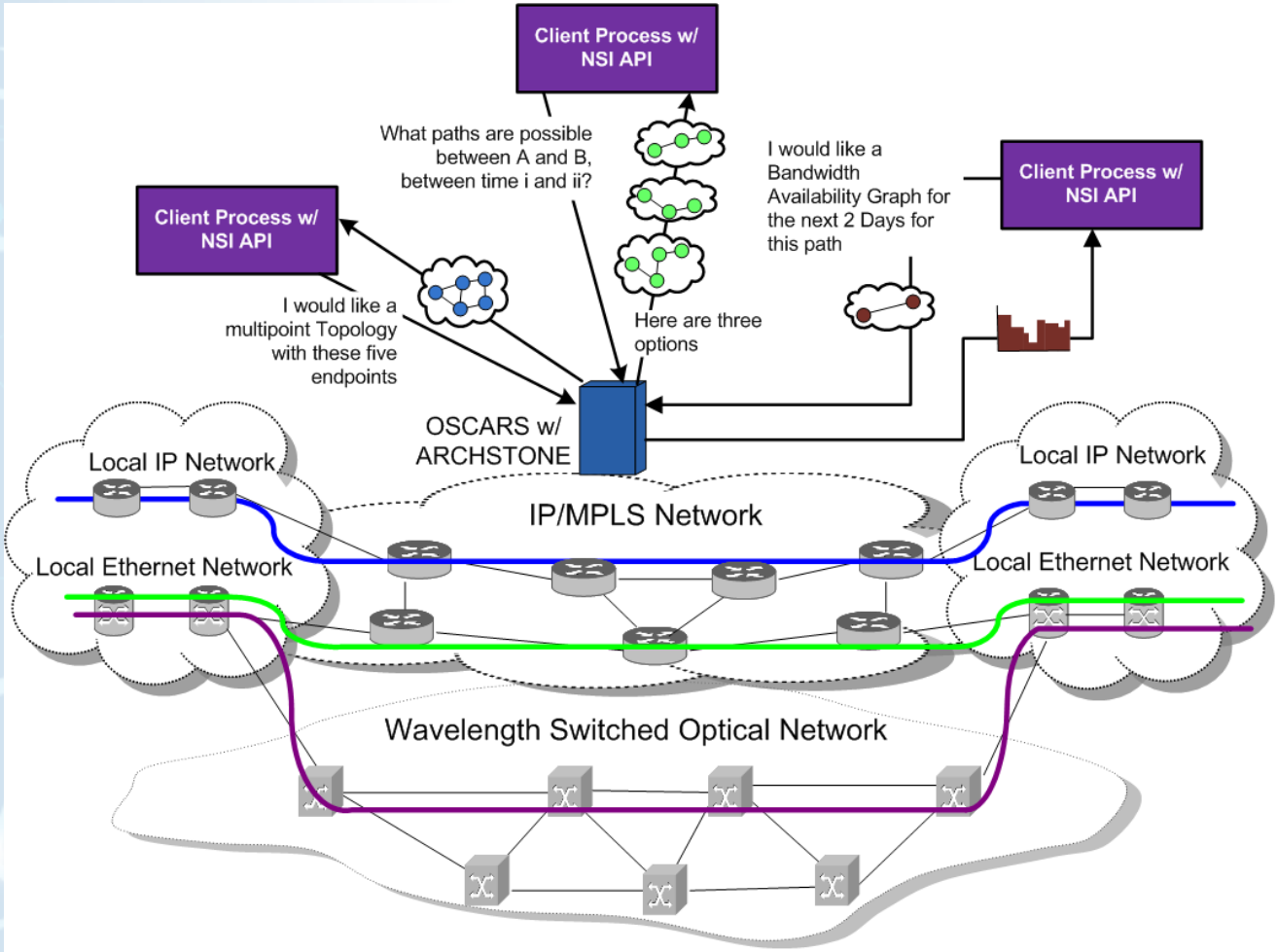
Science DMZ: remove roadblocks towards **end-to-end** performance

Science DMZ – a well-configured location for high-performance WAN-facing science services

- Located at or near site perimeter on dedicated infrastructure
- Dedicated, high-performance data movers
- Highly capable network devices (wire-speed, deep queues)
- Virtual circuit connectivity
- Security policy and enforcement are specific to science workflows
- perfSONAR



Programmable networks (1): Service Interfaces

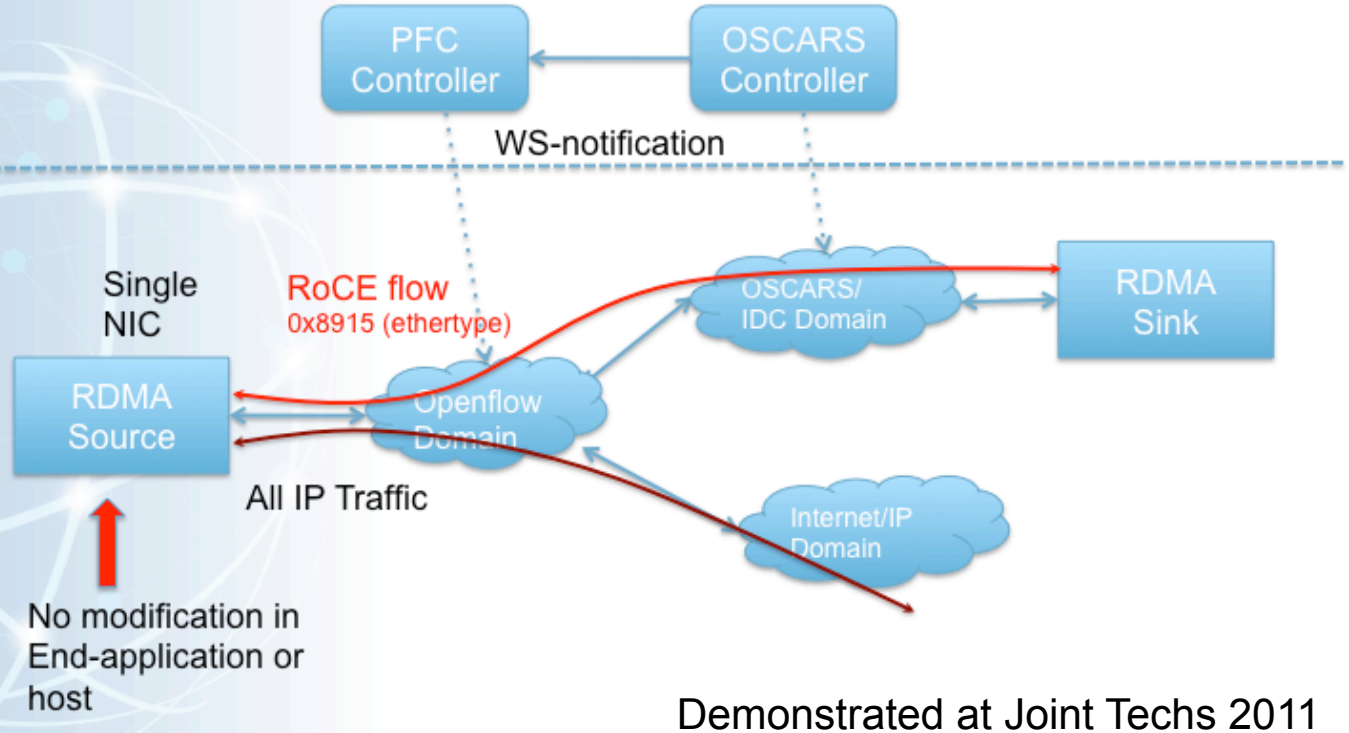


- Intelligent Network Services
  - Reservation
  - Scheduled
- Standard Service Interface
  - NSI (OGF)

Slide courtesy: ARCHSTONE project

Programmable networks (2): Software-Defined Networking (SDN)

Demonstrating end-to-end RDMA flows



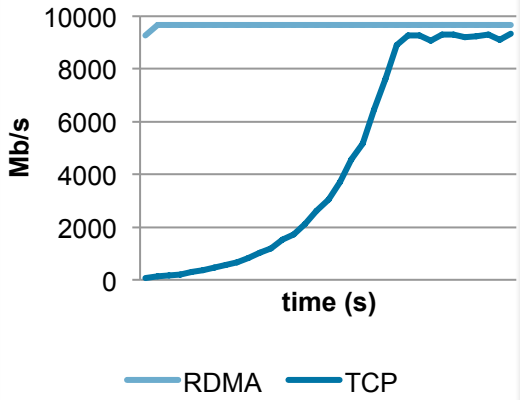
Demonstrated at Joint Techs 2011

Flexible, Programmable, separation of flows

Eric Pouyoul, Inder Monga, Brian Tierney (ESnet), Martin Swany (Indiana) & Ezra Kissel (U. of Delaware)

Programmable networks (3): New Protocols for high-speed data transfer

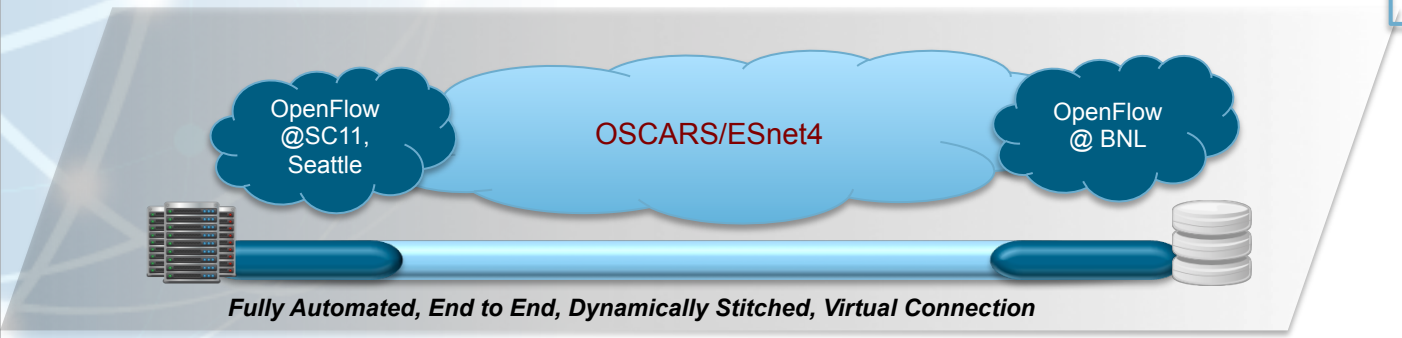
- Bridging end-site dynamic flows with WAN dynamic tunnels
  - Zero-configuration virtual circuit from end-host to end-host
  - Automated discovery of circuit end-points
- Cross-country RDMA-over-Ethernet high-performance data transfers
- SC11 Demonstration
  - 9.8 Gbps on 10 Gbps ,78 ms RTT link between Brookhaven in NY to Seattle, WA
  - <4% CPU load compared to 1 stream TCP w/80% util.
  - No special host hardware other than NIC with RoCE support



CPU core utilization

70-90% - 1 stream TCP

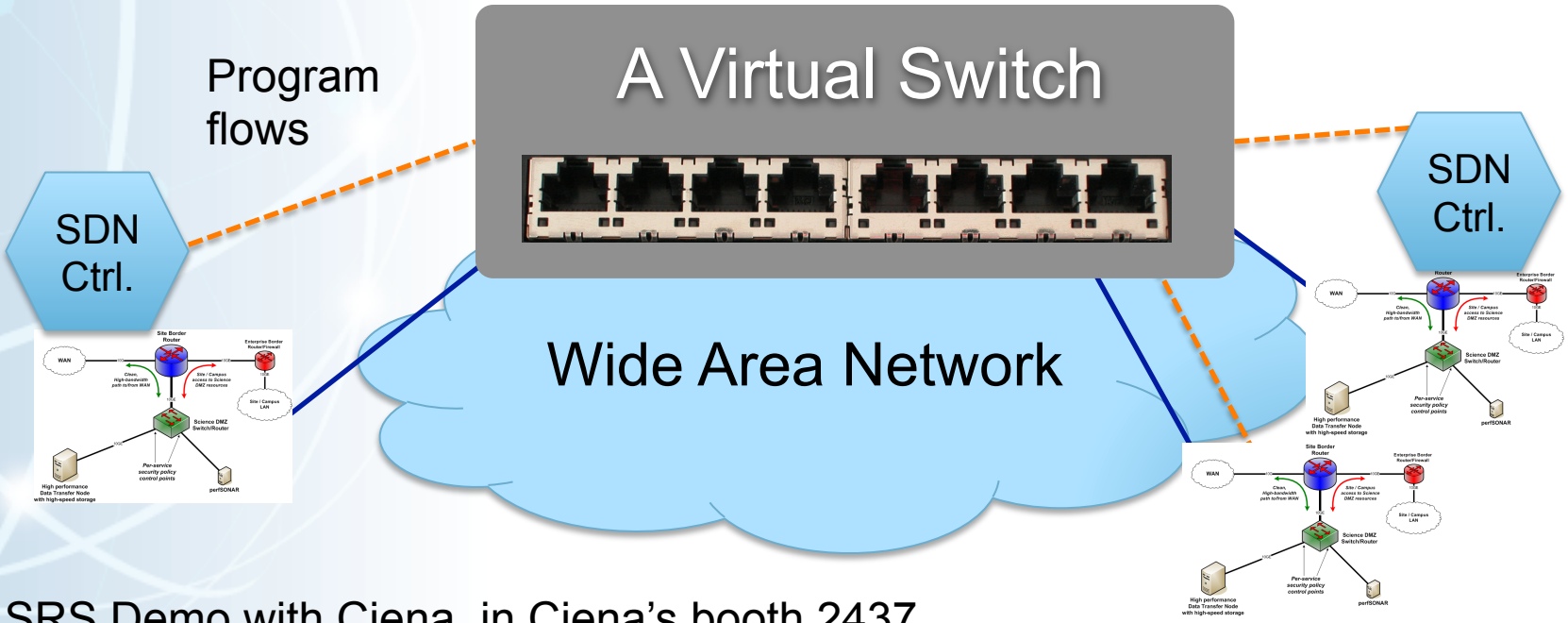
3-4% - RDMA



Fully Automated, End to End, Dynamically Stitched, Virtual Connection

Simple abstractions

How can Storage leverage this simple network abstraction?



SRS Demo with Ciena, in Ciena's booth 2437

Conclusion

Moving data fast(*er*) is a 21st century reality

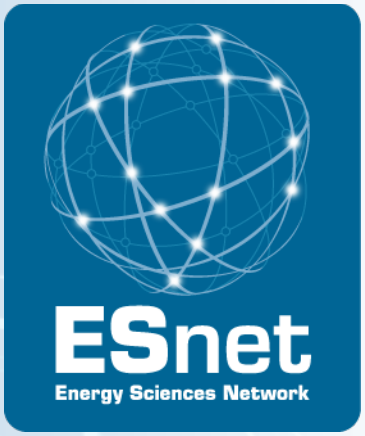
- Distributed Science Collaborations, Large Instruments, Cloud Computing

Network is not an infrastructure, but an instrument

- Think Different, do not set your expectations about the network as your traffic highway

Simple, **Programmable**, **Network abstractions** with a **Service Interface**

- How will storage workflows leverage that?



Inder Monga

imonga@es.net

www.es.net, fasterdata.es.net

Thank You!