# Distributed Big Data Assets Sharing & Processing

## Trusted Data Processing in Untrusted Environments.

C. de Laat (moderator), L. Gommans, R. Wilson

**System & Network Engineering, University of Amsterdam**
**AirFrance KLM**
**CIENA**

INTERNET2® 2017 global SUMMIT

April 23-26
Washington DC

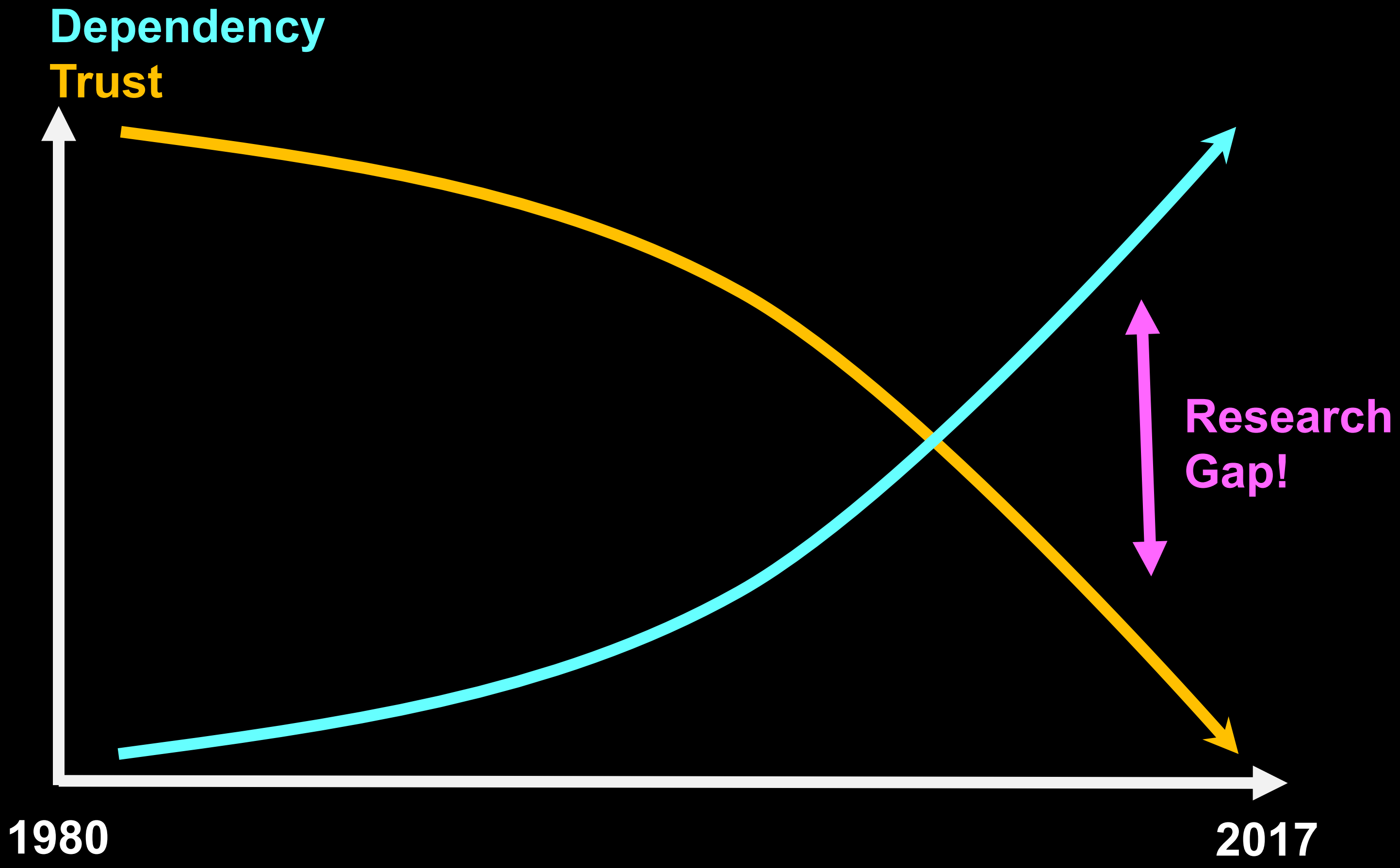
AIRFRANCE KLM

ciena

COMMIT/

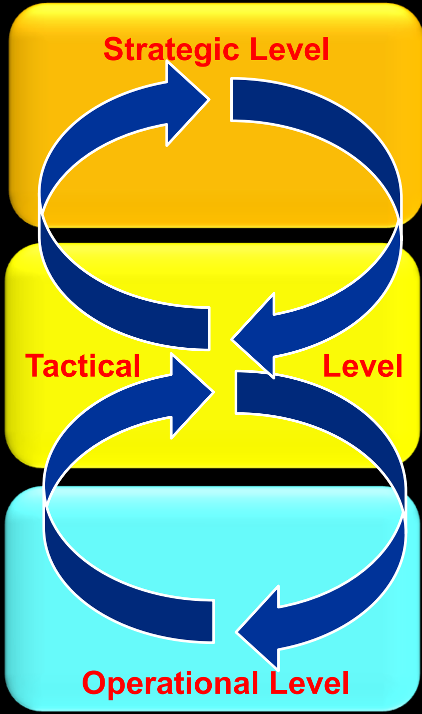
NWO
Netherlands Organisation
for Scientific Research

# Fading Trust in Internet

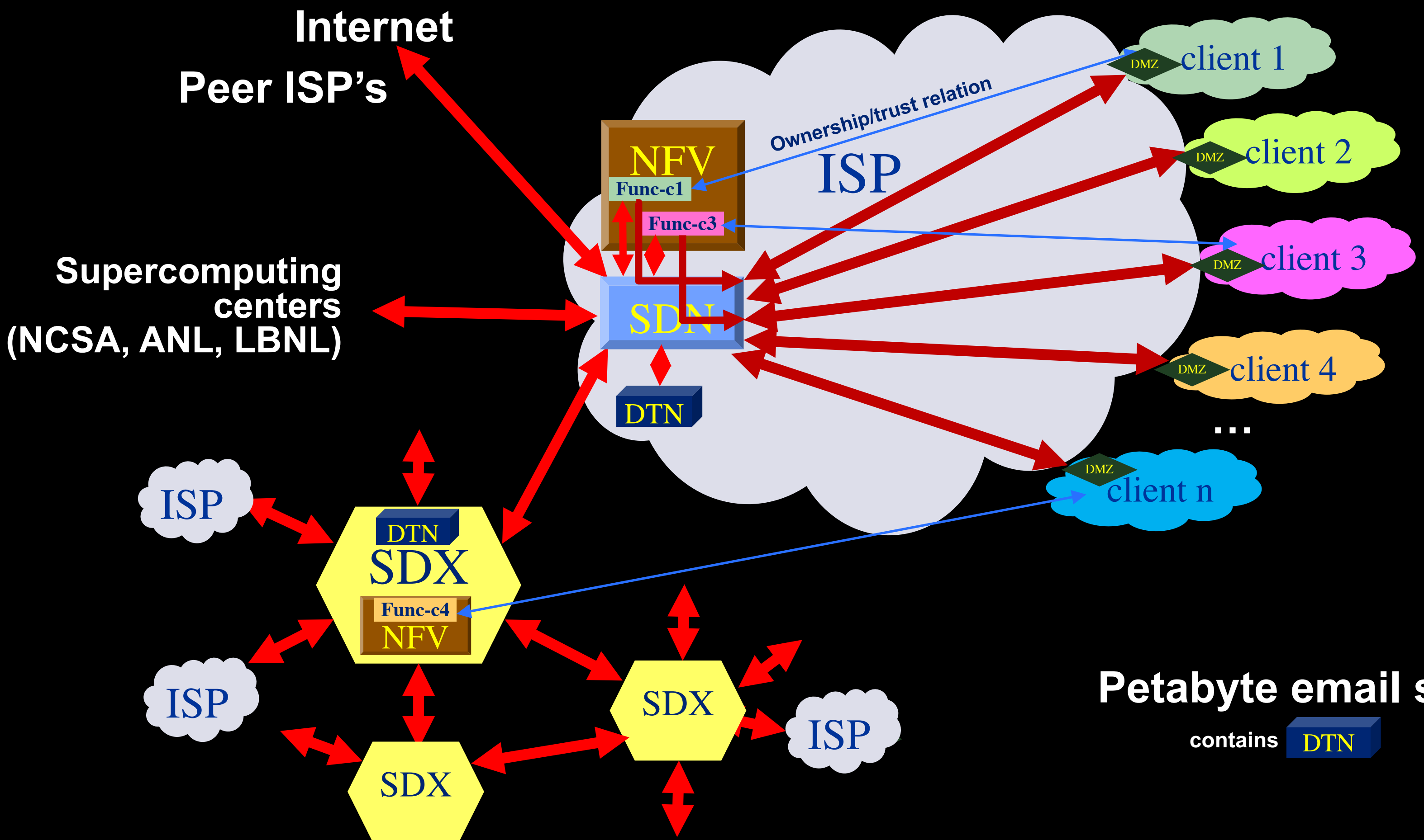


# Main problem statement

- Organizations that normally compete have to bring data together to achieve a common goal!
- The shared data may be used for that goal but not for any other!
- Data may have to be processed in untrusted data centers.
  - How to enforce that using modern Cyber Infrastructure?
  - How to organize such alliances?
  - How to translate from strategic via tactical to operational level?
  - What are the different fundamental data infrastructure models to consider?



# Networks of ScienceDMZ's & SDX's

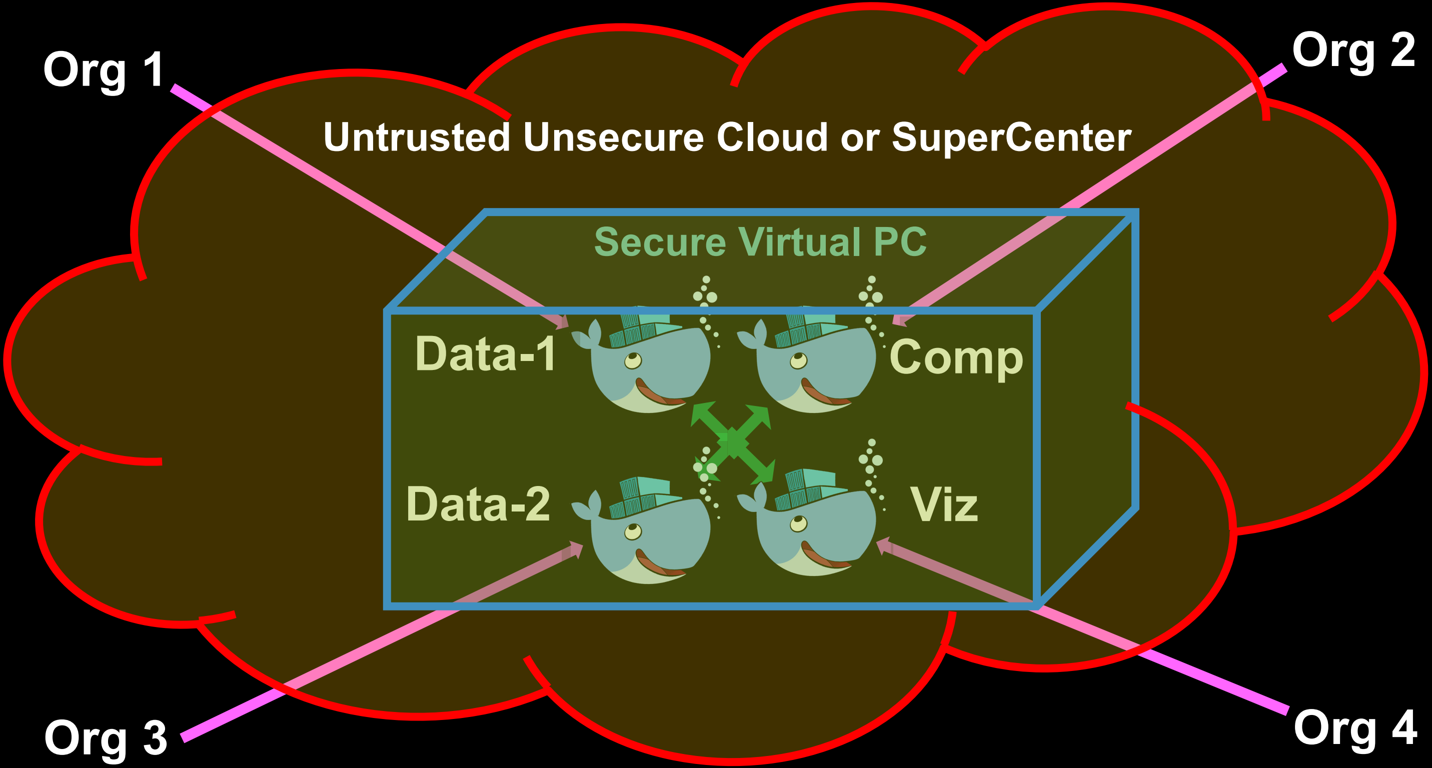


**Petabyte email service ☺**

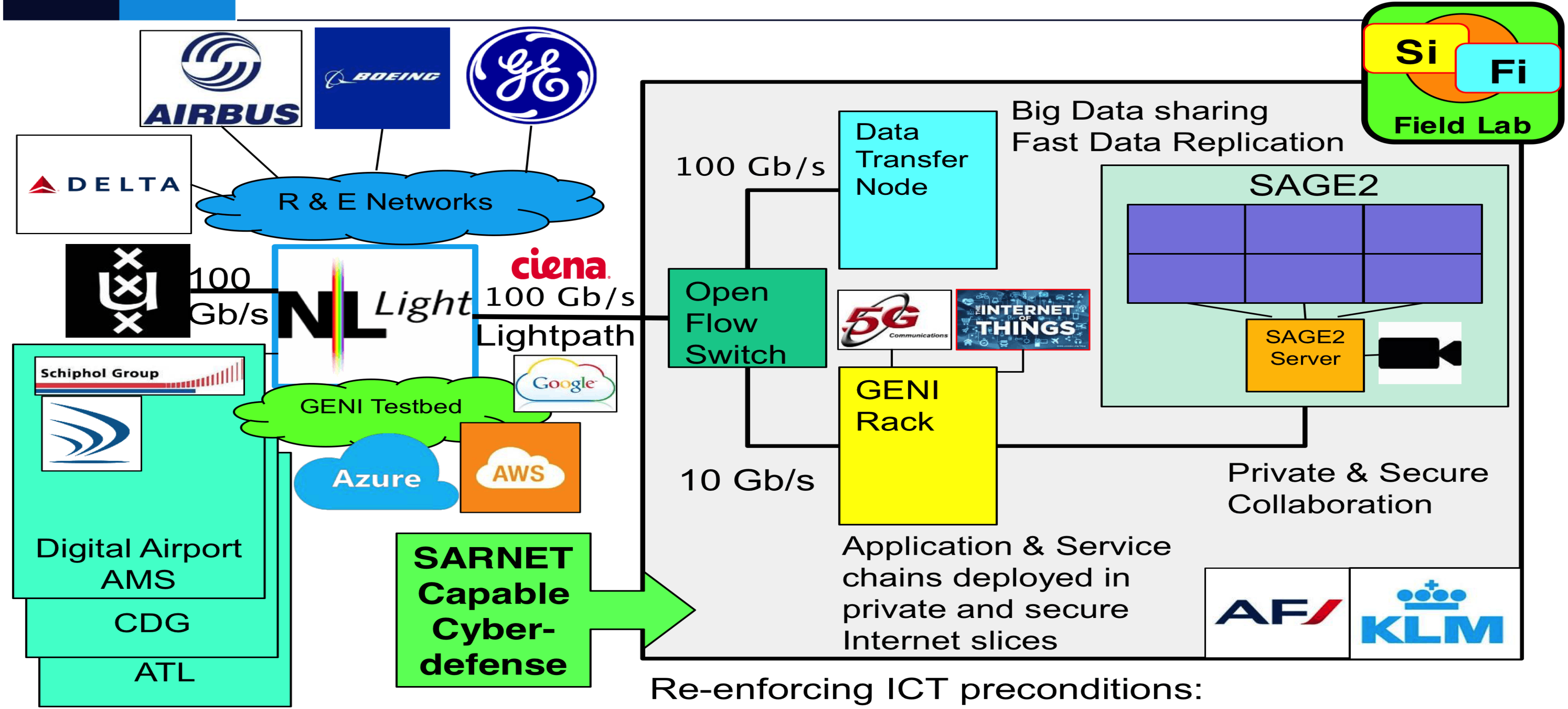
contains DTN

# Secure Policy Enforced Data Processing

- Bringing data and processing software from competing organisations together for common goal
- Docker with encryption, policy engine, certs/keys, blockchain and secure networking
- Data Docker (virtual encrypted hard drive)
- Compute Docker (protected application, signed algorithms)
- Visualization Docker (to visualize output)



# Ambition to put capabilities into fieldlab



Si Fi Field Lab

AIRBUS

BOEING

R & E Networks

DELTA

100 Gb/s

NL Light

ciena

100 Gb/s

Lightpath

Google

GENI Testbed

Azure

AWS

Schiphol Group

Digital Airport AMS

CDG

ATL

SARNET Capable Cyber-defense

100 Gb/s

Data Transfer Node

Big Data sharing
Fast Data Replication

SAGE2

SAGE2 Server

Open Flow Switch

5G Communications

THE INTERNET OF THINGS

GENI Rack

10 Gb/s

Private & Secure Collaboration

Application & Service chains deployed in private and secure Internet slices

AF KLM

Re-enforcing ICT preconditions:
Each envisaged site has similar elements

**Program**:

15h00 Cees de Laat, University of Amsterdam

Trusted Data Processing in Untrusted Environments.

15h05 Leon Gommans, Air France KLM

Trusted Big Data Sharing.

15h25 Rodney Wilson

Programmable Supernetworks, Science DMZ based Networking.

15h30 Panel of stakeholders Flash talks (~3 min each):

Inder Monga - ESnet - Data Science Driving Discovery.

Matt Zekauskas - Internet2 - Thoughts on Internet2 and Trusted Large Data Transfer.

Jerry Sobieski - NORDUnet - Issues of Big Data Sharing in a Global Science Collaboration.

Adam Slagell – NCSA - What are we trusting?

15h45 Panel discussion moderated by Cees de Laat

16h00 End of session.

INTERNET2®

2017 global SUMMIT

# TRUSTED BIG DATA ASSET SHARING

**Leon Gommans**

Science Officer, Air France KLM Group IT Technology Office

Guest Researcher, University of Amsterdam

AIR FRANCE KLM

Leibniz Center for Law

SNE System and Network Engineering

BRINGING
NETWORKS
TOGETHER

gs17

## CONTENT

- Sharing Big Data Assets and Trust
- Secure Digital Market Place concept
- Infrastructure model research
- Research project involvement.

# Sharing Big Data Assets within a group needs

Clearly defined and agreed common **benefit** defining the group's identity

**Common group rules** governing use, access and benefit sharing.

**Organizing trust** amongst group members as **means to reduce risk**

Infrastructure supporting **implementation of trust** whilst ensuring **autonomy**

# Trust as a means to reduce risk

Risk:

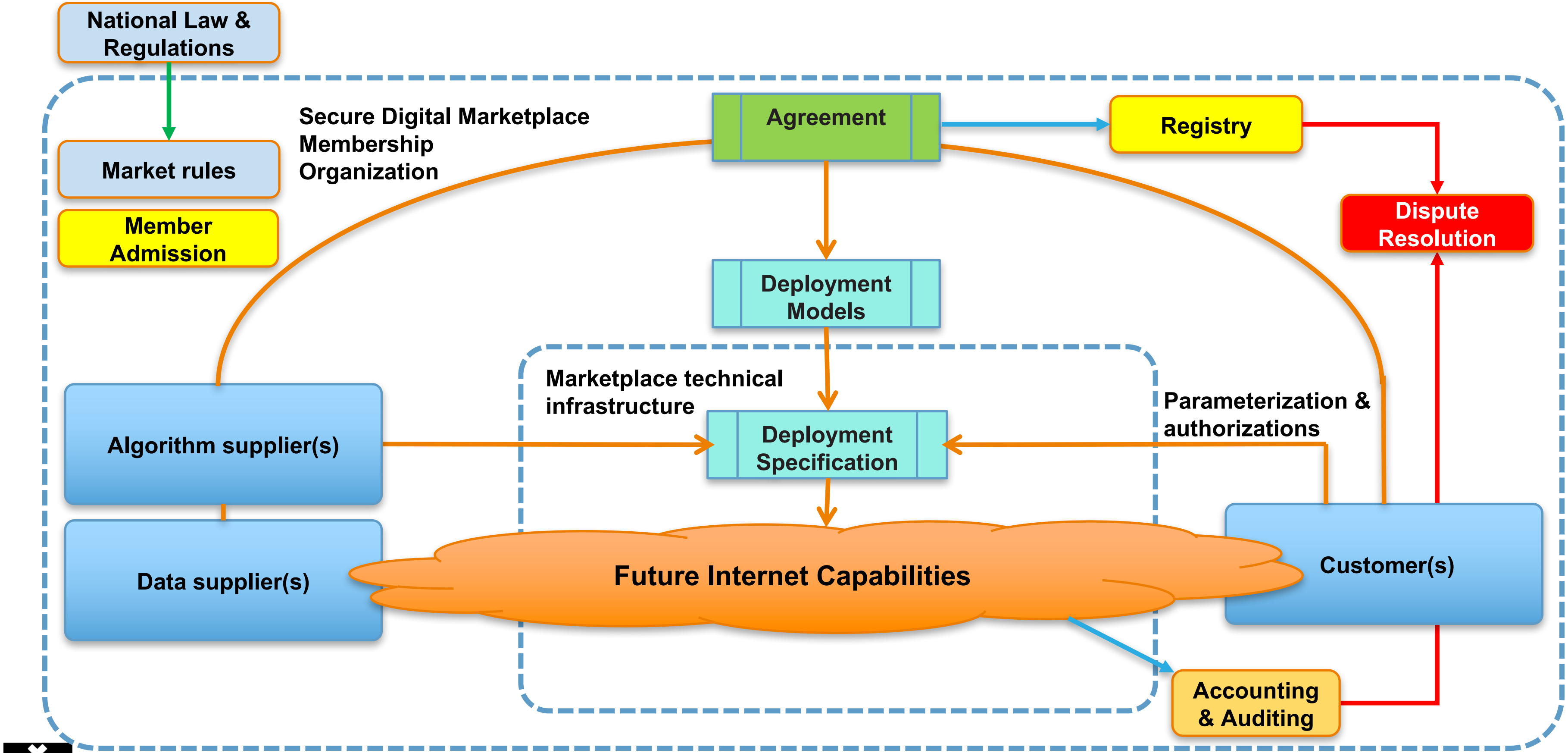
Compliancy
Liability
Disclosure
Ownership
Intellectual Property
Additional oversight
etc., etc...

Means:

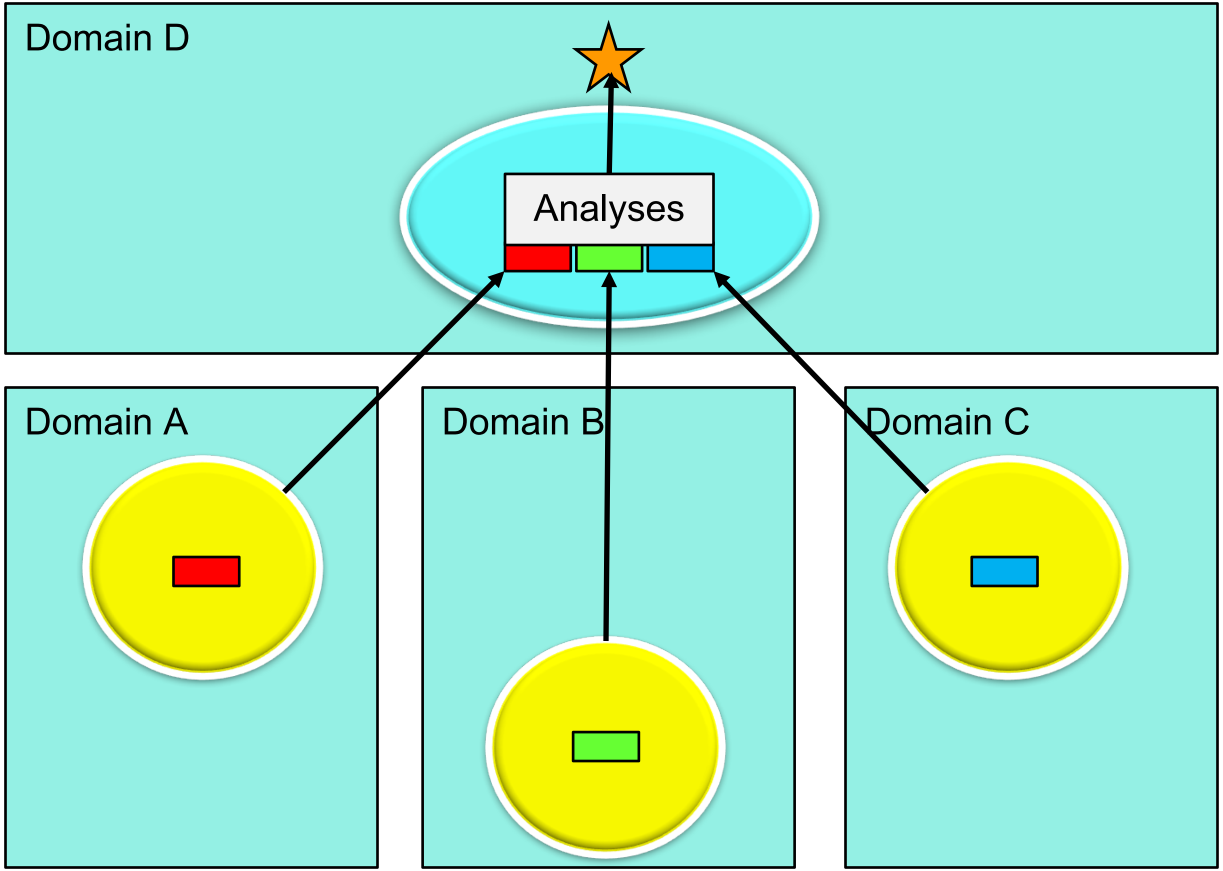
Trust and power are both means capable of reducing risk

How to organize trust and power? -> **The Secure Digital Market Place concept**

# The Secure Digital Market Place: A high level framework

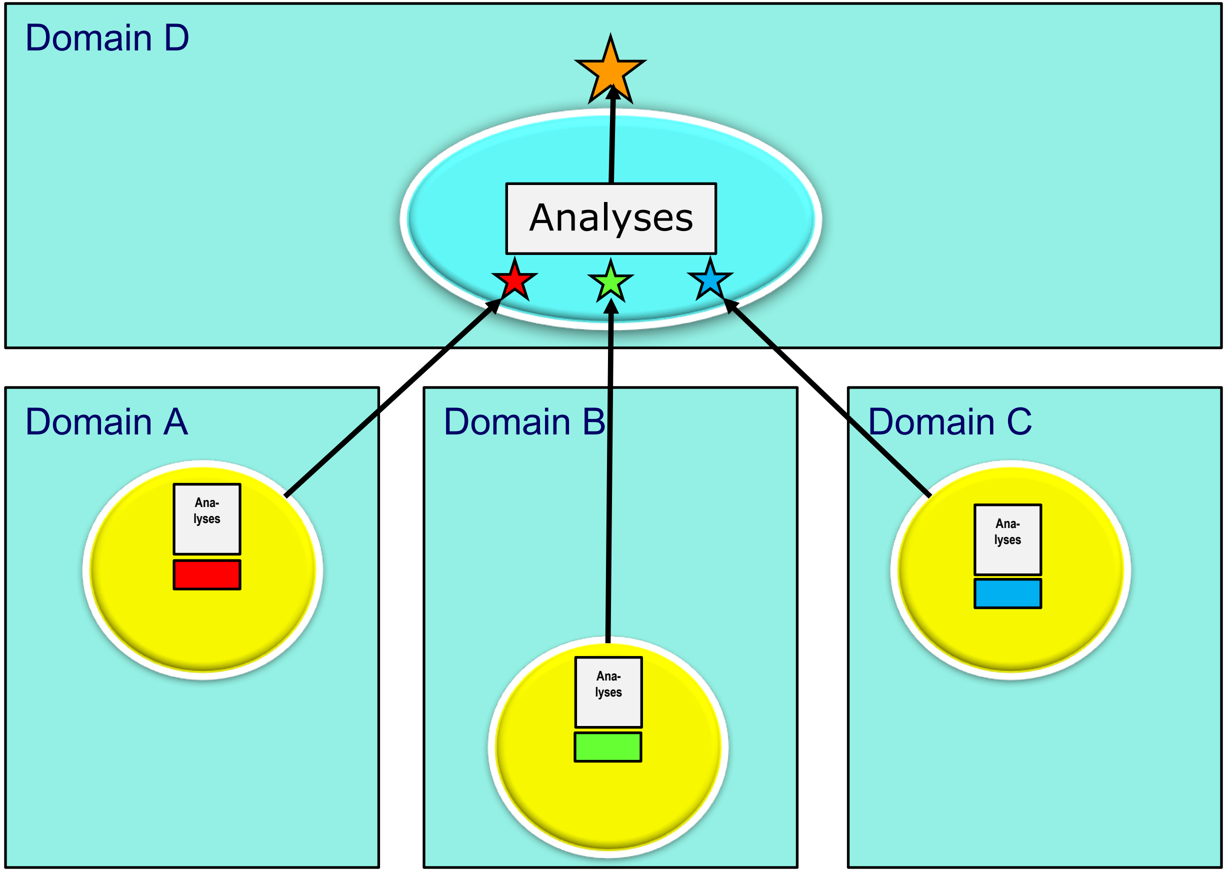

National Law & Regulations
Market rules
Member Admission
Secure Digital Marketplace Membership Organization
Agreement
Registry
Dispute Resolution
Deployment Models
Marketplace technical infrastructure
Parameterization & authorizations
Algorithm supplier(s)
Deployment Specification
Data supplier(s)
Future Internet Capabilities
Customer(s)
Accounting & Auditing

# Traditional Model raising concerns

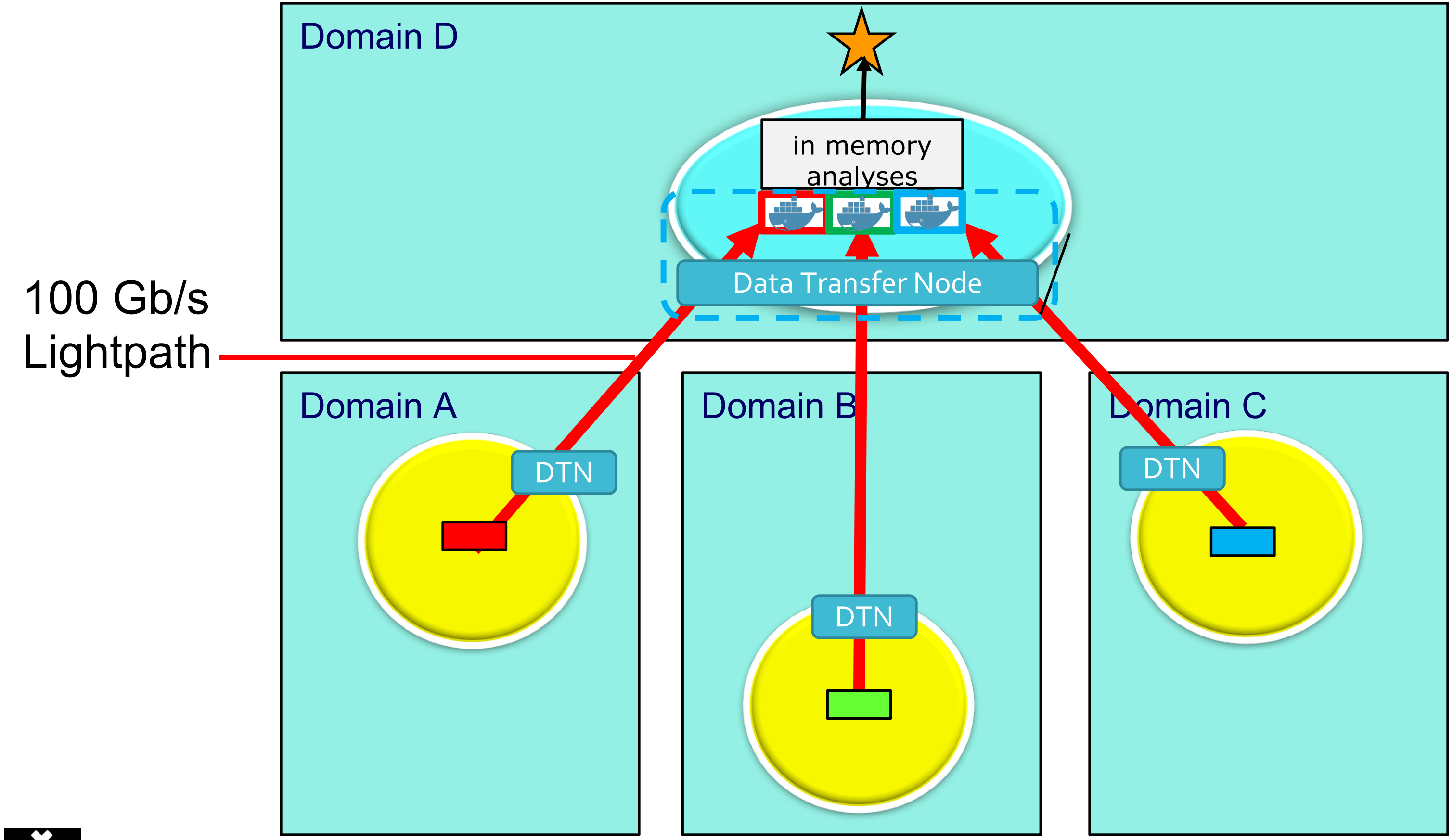


Domain = Autonomous Organization with own administration and enforcement

# Alternative: bring processing to the data



# An innovative deployment model: separate processing from data



**Data Transfer Node** enables utilization of available high network bandwidth across distance

DTN is part of Science DMZ concept from

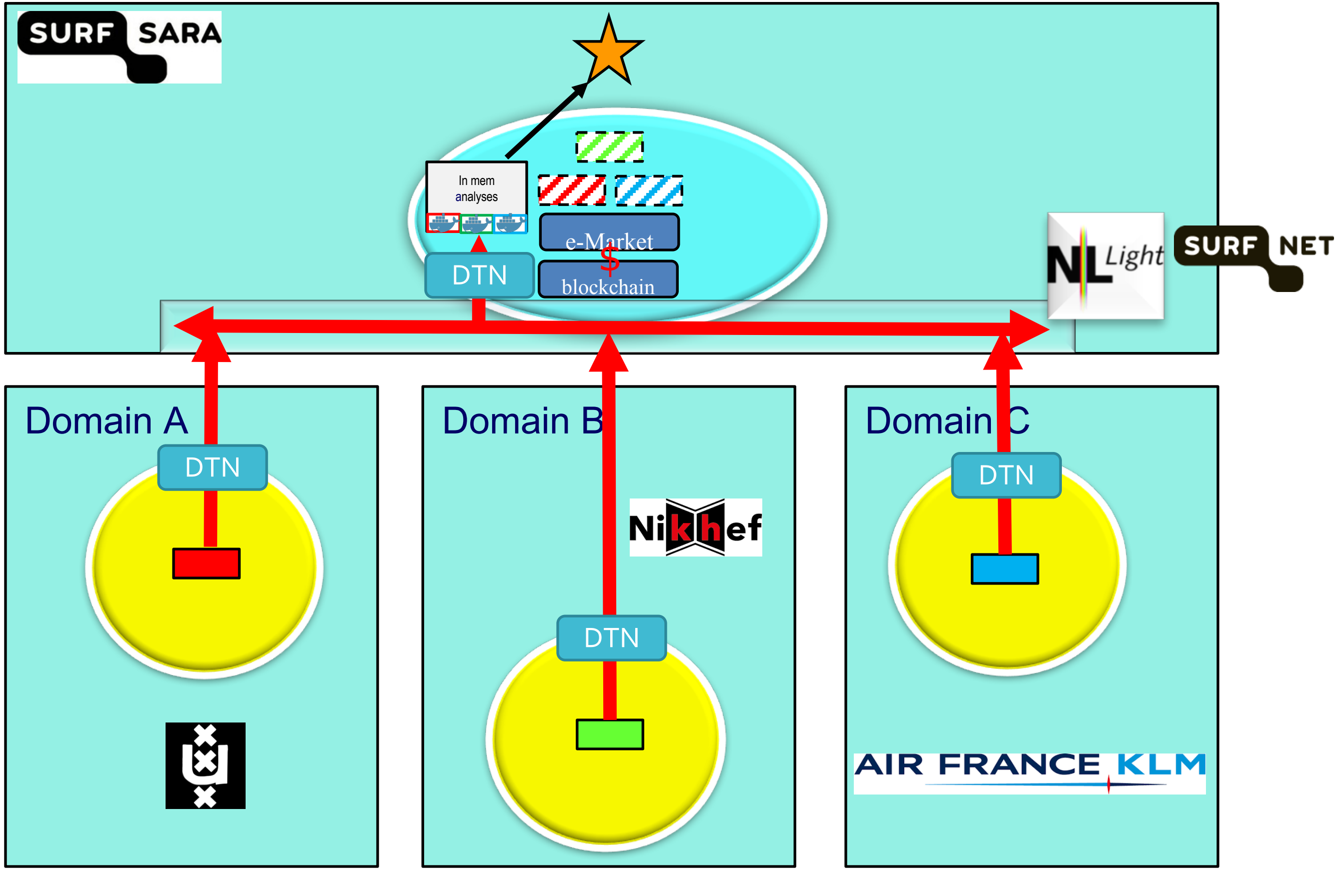
ESnet ENERGY SCIENCES NETWORK

PRP PACIFIC RESEARCH PLATFORM

System and Network Engineering

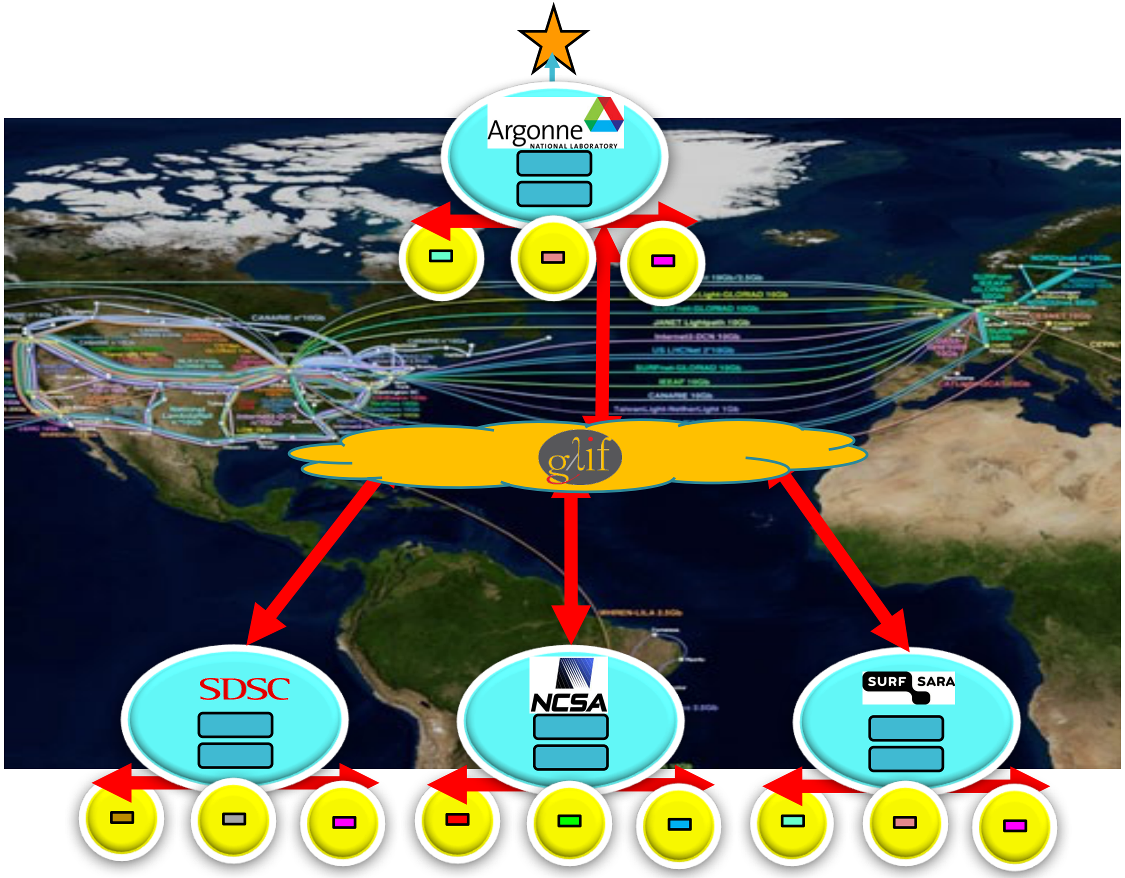
AIR FRANCE KLM

# Secure Digital Market Place deployment model research testbed



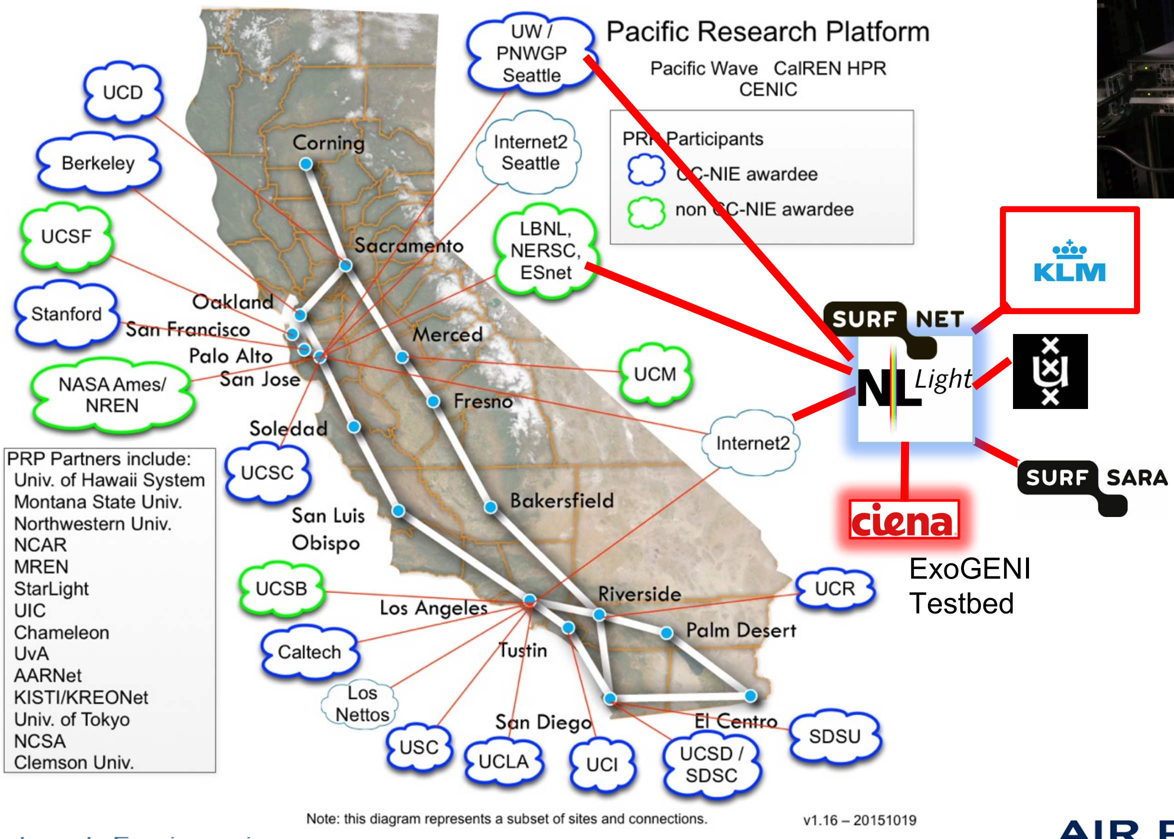
System and Network Engineering

# Global Digital Market Place Testbed via the GLIF?



# Pacific Research Platform testbed involvement

**Research goal:**
Explore value of academic network research capabilities that enable innovative ways & models to share big data assets



PRP Partners include:
- Univ. of Hawaii System
- Montana State Univ.
- Northwestern Univ.
- NCAR
- MREN
- StarLight
- UIC
- Chameleon
- UvA
- AARNet
- KISTI/KREONet
- Univ. of Tokyo
- NCSA
- Clemson Univ.

Note: this diagram represents a subset of sites and connections.

v1.16 – 20151019

ExoGENI Testbed

Data Transfer Node at KLM fieldlab with 100 gb/s link to enable SDMP research thanks to UvA, SURFnet and Ciena

# Big Data Sharing use cases placed in airline context

**Global Scale**

**National Scale**

**City / regional Scale**

**Campus / Enterprise Scale**

Aircraft Component Health Monitoring (Big) Data
NWO **CIMPLO project**
4.5 FTE

Cargo Logistics Data
**NLIP iShare project**

iSHARE
powered by NLIP

Cybersecurity Big Data
NWO COMMIT/
**SARNET project**
3.5 FTE

AIR FRANCE KLM

SNE System and Network Engineering

INTERNET2®

2017 global SUMMIT

# Thank you !

NL Research funded by **NWO, STW, COMMIT/, Commit2Data, NLIP**
in collaboration with **Internet2, ESnet, PRP, NCSA, ANL, ICAIR,..**

**University of Amsterdam:** Cees de Laat, Tom van Engers, Paola Grosso, Ameneh Deljoo, Gleb Polevoy, Ralph Koning, Ben de Graaff, Lukasz Makovski
**Ciena:** Steve Alexander, Rodney Wilson, Marc Lyonais, Lance Williford
**SURFnet:** Erik Huizer, Gerben van Malenstijn
**SURFsara:** Anwar Osseyran, Axel Berg
**Leiden University:** Thomas Baeck, Jeroen van der Leijé
**TNO:** Rob Meijer, Frank Franssen, Jan Burgmeijer, Jan Wester
**CWI:** Marc Stevens
**Air France KLM:** Edwin Borst, Nicolas Forgues, Vincent Euzeby, Bart Krol, Wouter Kalfsbeek
**NLIP / iShare** Michiel Haarman, Vincent Janssen, Gijs Burgers

BRINGING
NETWORKS
TOGETHER

gs 17

ciena
Experience. Outcomes.

INTERNET2 2017 global SUMMIT
April 23-26
Washington DC

# Programmable Supernetworks, Science DMZ based Networking

Rodney G. Wilson
Sr. Director, External Research Programs
CTO - Ciena

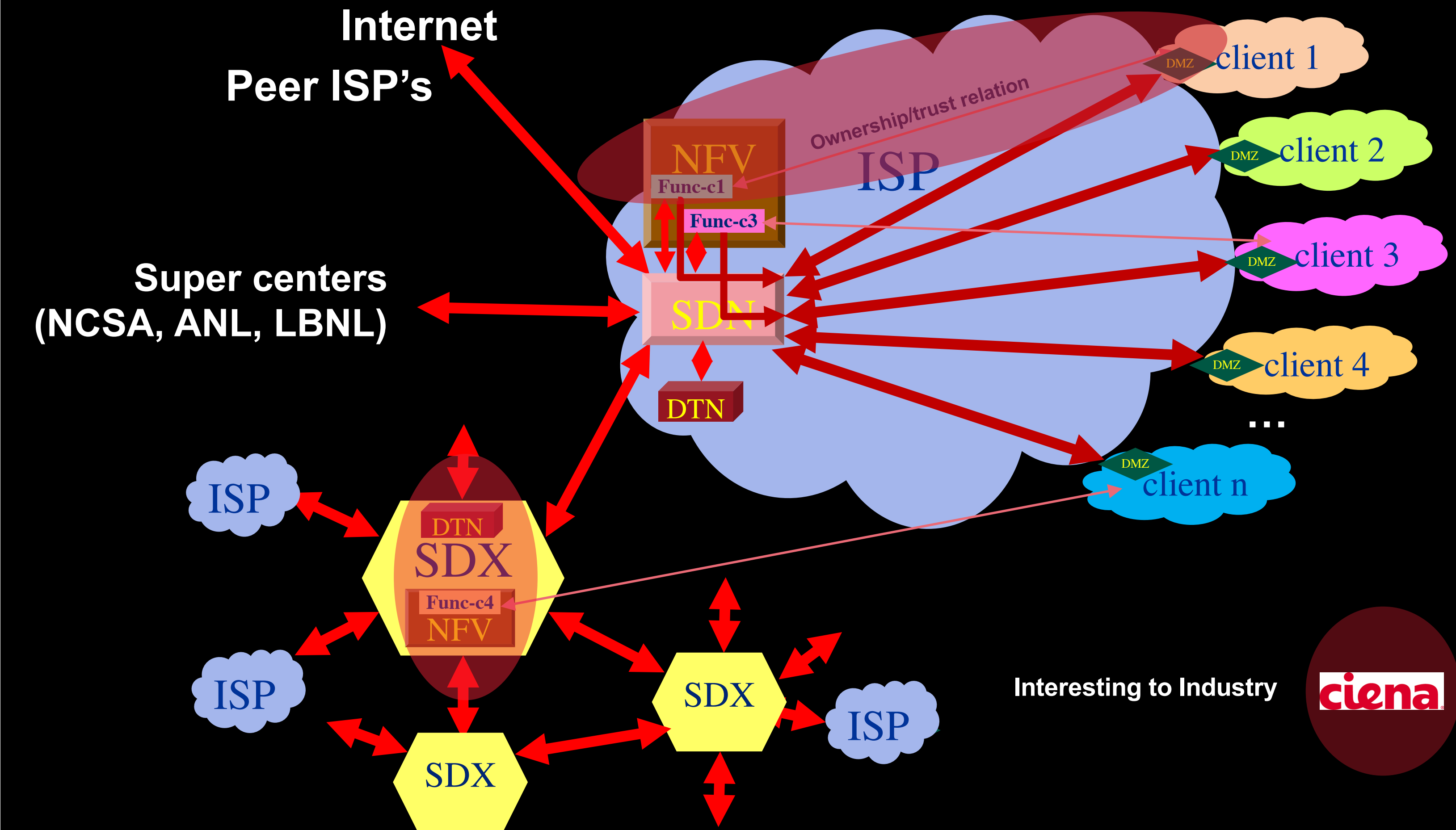
# Industry Interests

Issues in moving large dataflows

We have issues with trust & security

Tomorrow’s problems today

# Putting theory in to practice…



Internet

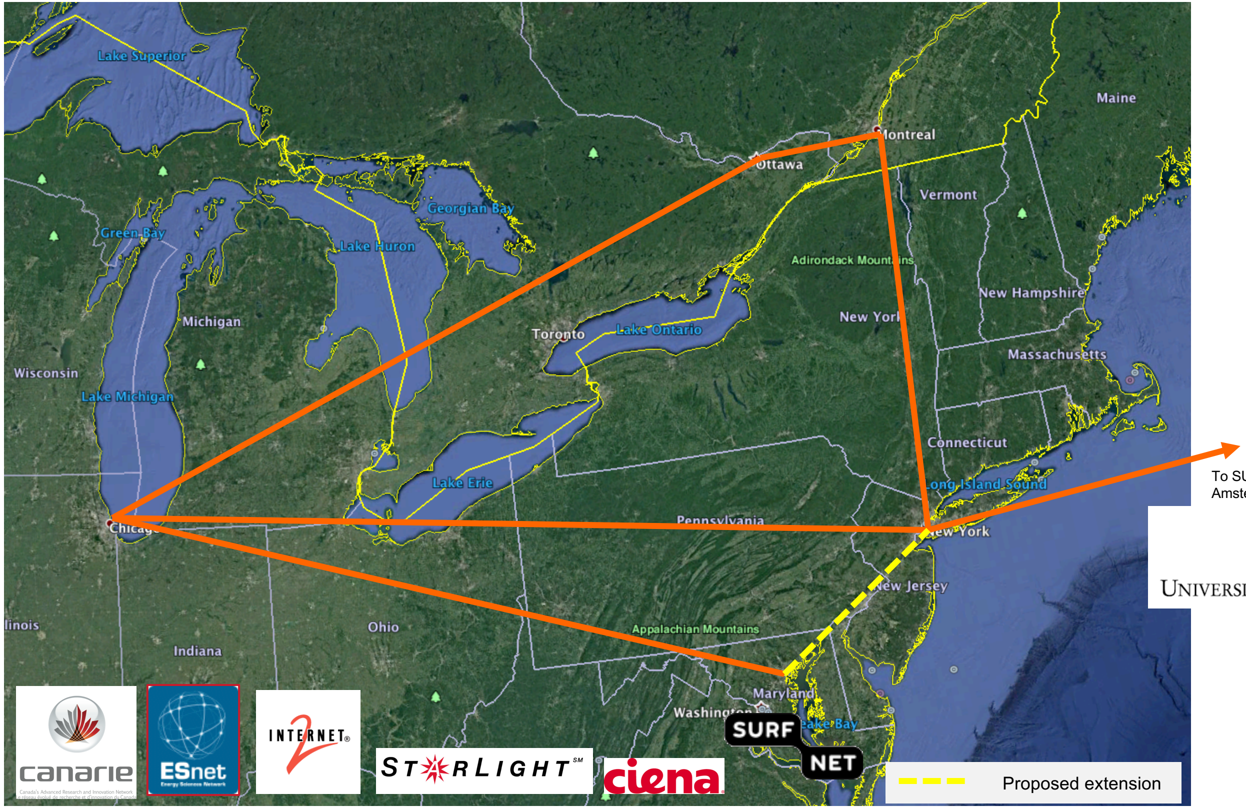
Peer ISP’s

Super centers
(NCSA, ANL, LBNL)

Interesting to Industry

ciena

# Field lab



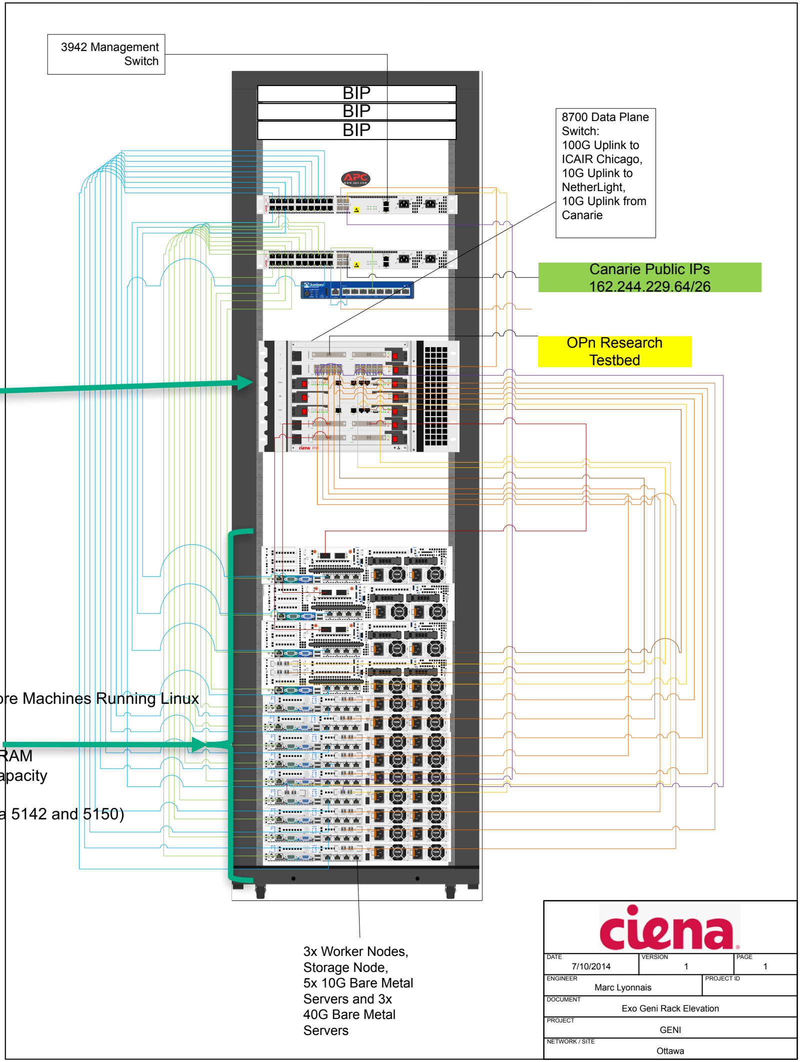
To SURFnet
Amsterdam

UNIVERSITY OF AMSTERDAM

# CENI “client resource”
# NFV Engines
# DMZ vs. lockdown

**8700 Packet Wave Platform**

- ❑ 4 Slot with 560G of L2 Capacity
- ❑4x40G ( 2 PSLM-200-2)
- ❑2x100G ( 1 PSLM-200-2)
- ❑20x10GE (1 PSLM-200-20)

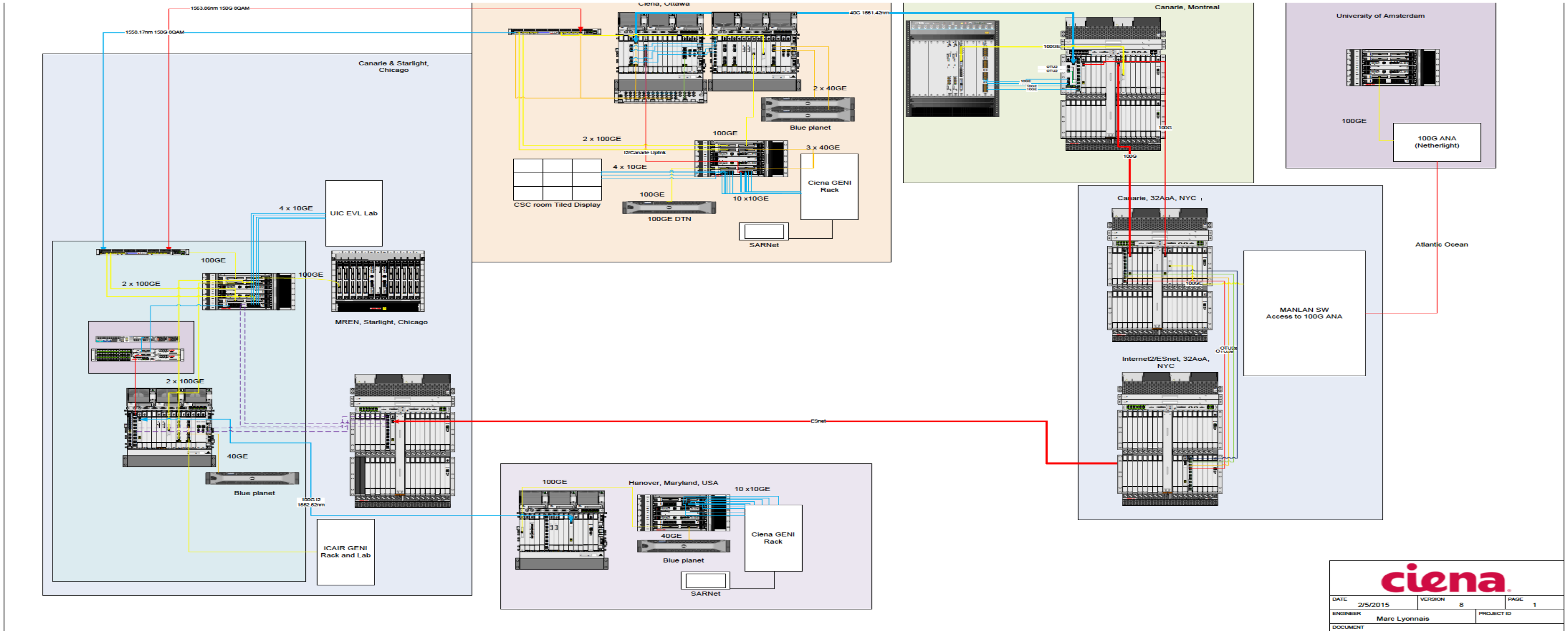


| | | |
|---|---|---|
| DATE 7/10/2014 | VERSION 1 | PAGE 1 |
| ENGINEER Marc Lyonnais | PROJECT ID | |
| DOCUMENT Exo Geni Rack Elevation | | |
| PROJECT GENI | | |
| NETWORK / SITE Ottawa | | |

**CENI Ottawa System Specifications**

- ❑14 Dell Servers
  - ❑180 Physical Cores -> approx. 330 Virtual Core Machines Running Linux RedHAT 6.0
  - ❑Up to ~ 80 VMs (using 4 Cores each.)
  - ❑608 GB of Physical RAM -> approx. 1.2TB VRAM
  - ❑6 TB of HD-> more than 12TB Virtual Disk Capacity
- ❑100GE Upload Capacity, first of its kind for GENI
- ❑20GE in Management Ethernets ports (approx 48 ports) via 5142 and 5150)
- ❑All DC powered ( approx. 100A)
- ❑175 Public IP addresses on CANARIE Network

# Field Lab Architecture



DATE 2/5/2015
VERSION 8
PAGE 1
ENGINEER Marc Lyonnais
PROJECT ID
DOCUMENT

# Data Science Driving Discovery

ESnet

**Inder Monga**

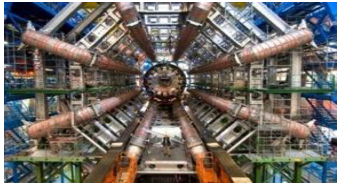
Director, ESnet

Division Director, Scientific Networking

Lawrence Berkeley National Lab

# Rapid increase in data rates and number of data sources

large experimental facilities …



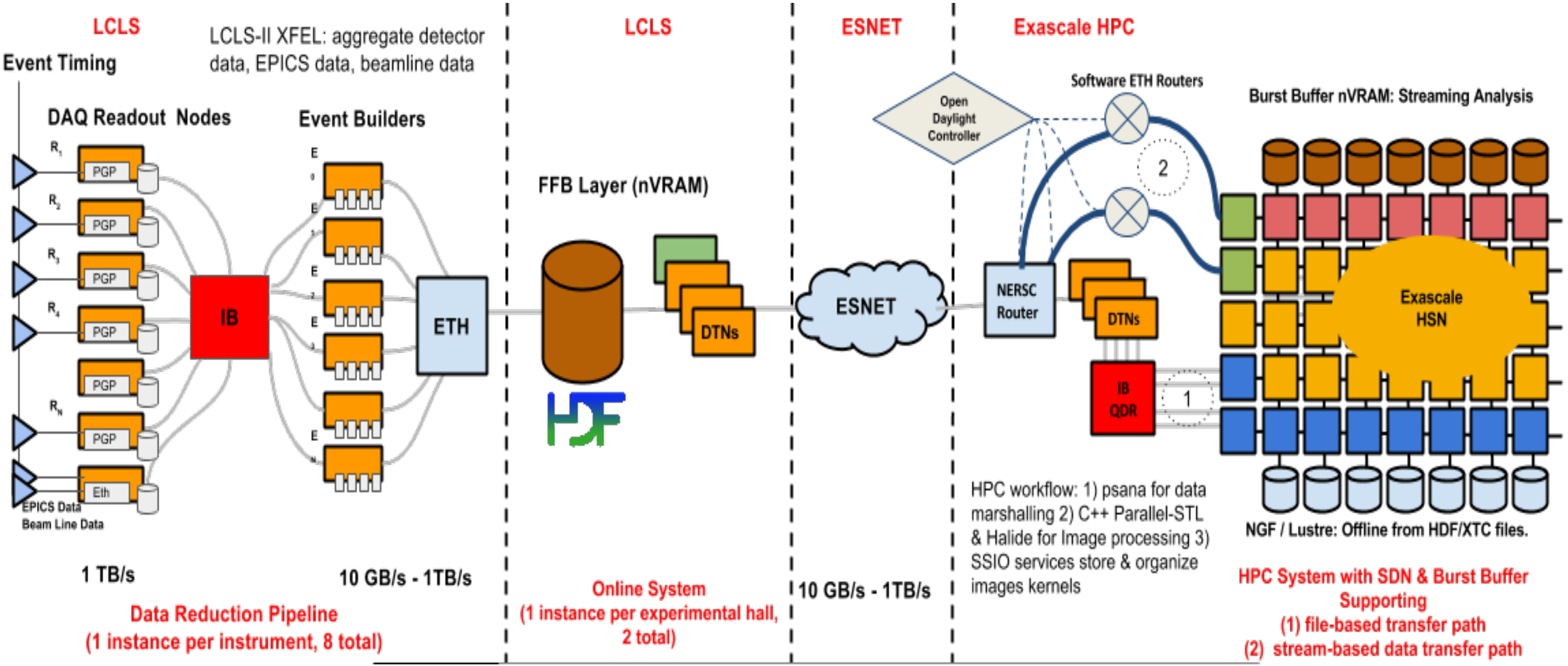
- Complexity of scientific discovery increasing
- Data volumes are increasing > Moore's Law
- Fewer large facilities, but global scientific population

**Automated coupling of compute and storage with networks critical to increasing science productivity**

..and smaller, new data sources.

# LCLS to NERSC – ECP ExaFEL project

Providing atomic scale vision to researchers at beamline in < 10min



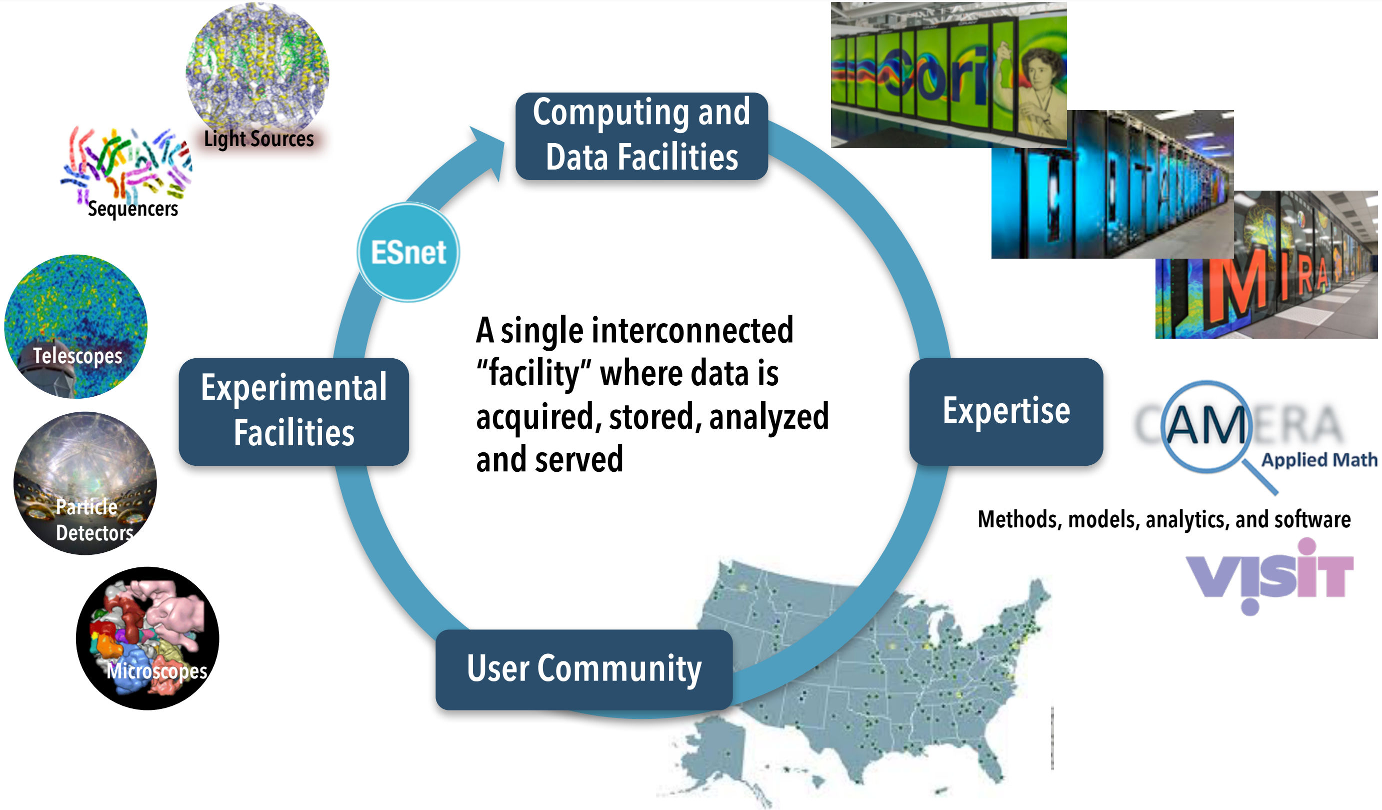
Data Reduction Pipeline
(1 instance per instrument, 8 total)

Online System
(1 instance per experimental hall, 2 total)

HPC workflow: 1) psana for data marshalling 2) C++ Parallel-STL & Halide for Image processing 3) SSIO services store & organize images kernels

HPC System with SDN & Burst Buffer Supporting
(1) file-based transfer path
(2) stream-based data transfer path

# Superfacility: Integrated network of experimental and computational facilities and expertise



Light Sources

Sequencers

Telescopes

Particle Detectors

Microscopes

Computing and Data Facilities

ESnet

Experimental Facilities

Expertise

User Community

A single interconnected "facility" where data is acquired, stored, analyzed and served

CAMERA Applied Math

Methods, models, analytics, and software

VisIt

INTERNET2

2017 global SUMMIT

BRINGING NETWORKS TOGETHER

gs 17

## Thoughts on Internet2 – Big Data Panel

**Matt Zekauskas**
matt@internet2.edu

# Thoughts on Internet2 and Trusted Large Data Transfer

- Internet2 builds a network to support these sort of big data transfers, connecting our regional networks, schools and service providers
- We can build custom paths, dynamically, to support communication among trusted partners
- The Internet2 community has also worked in trust and identity, creating the inCommon trust fabric, and the TIER program.  Leverage this to help create bilateral trust between entities
- Internet2 is involved with the Pacific Research Platform partners toward creating a national research platform, including “standards” for data transfer nodes – an opportunity to improve trusted big data flows
  - A way to collaboratively negotiate and articulate trust and thus access
  - Blend policy and social to reduce friction to discover and negotiate
- Increase transparency – telemetry – to foster trust?

# Issues of Big Data Sharing in a Global Science Collaboration

## Is it networking issue?
## Or is it a security issue?

Jerry  Sobieski
Chief Research Officer
NORDUnet

Presented to the Internet2 Global Summit 2017

# Sharing large data assets

- Redistributing and correlating large data has two major challenges:
  - Moving large data sets across large physical distances -> The classic network capacity/performance issue  (This assumes the two locations are trusted)
  - Secured access to information – once outside a secure perimeter, there is no longer effective control of access to that info.  (i.e. how do we “trust” remote locations?)
- Moving the algorithm to the data:
  - Useful where the distributed data sets are already integrated in a single “location”
  - Does not solve the problem of gathering distributed data sets for correlation or other integrated analysis algorithms,
- Exposes the algorithm to potential security breaches
  - Proprietary algorithms may be compromised

NORDUnet
Nordic infrastructure for Research & Education

# Security Issues in data sharing

- Jurisdictional restrictions
  - (E.g. national borders )
- Proprietary restrictions
  - e.g. business policy, IP algorithms
- Privacy restrictions
  - E.g. personal financial info, medical data, etc.
- Trust – but verify
  - Verifiably compliance – can we authorize each access of information? Or limit the use to a single trusted agent?
- Provinence – how do we handle provinence / reproducibility where data access is secured or constrained?

# The virtualized world

- “Virtualization” poses important challenges
  - The physical location of information is no longer determined
  - What constitutes a secure (trusted) perimeter in virtual service environments?
- “Cloud” services have not solved the security problem:
  - We can store encrypted data
  - We can transport encrypted data
  - We cannot [yet?] compute on encrypted data (homomorphic computing)
  - This exposes data in the clear

# Key open questions:

- Can we *verifiably* secure computational processes short of physical secure perimeters?
  - Security thru obscurity?  Distributed computation, interchangable algorithmic components,
  - Who verifies and signs "trusted" code – can we trust them? -> trusted security services who's business value proposition is their reliability in terms of security analysis of components.
  - Homomorphic (encrypted) computing?
- We can authorize access to information, but having authorized access to some agent, we lose control over the information because that info is now in the clear...
  - Can we encrypt and "sign" data in such a way that only authorized agent(s) can interpret the data and make use of it?

# NATIONAL CENTER FOR SUPERCOMPUTING APPLICATIONS

2017 Internet2 Global Summit *Panel on Big Data Sharing*

Adam Slagell

# What are we trusting?

- Trusting not to re-share?
- Trusting to act competently?
- Trusting our common incentives and aligned interests?
- Trusting algorithms and expertise?
  - Requires deep understanding of the problem, right semantics for policies
  - Does not often generalize

# What needs to be shared?

- Analysis can happen at many layers
  - REN-ISAC members share derived data, not raw
- Can we do non-consumptive analysis?
  - Depends on the problem space
- Sharing publicly very hard
  - Understand the limits of anonymization
- Start simple
  - One well-defined problem