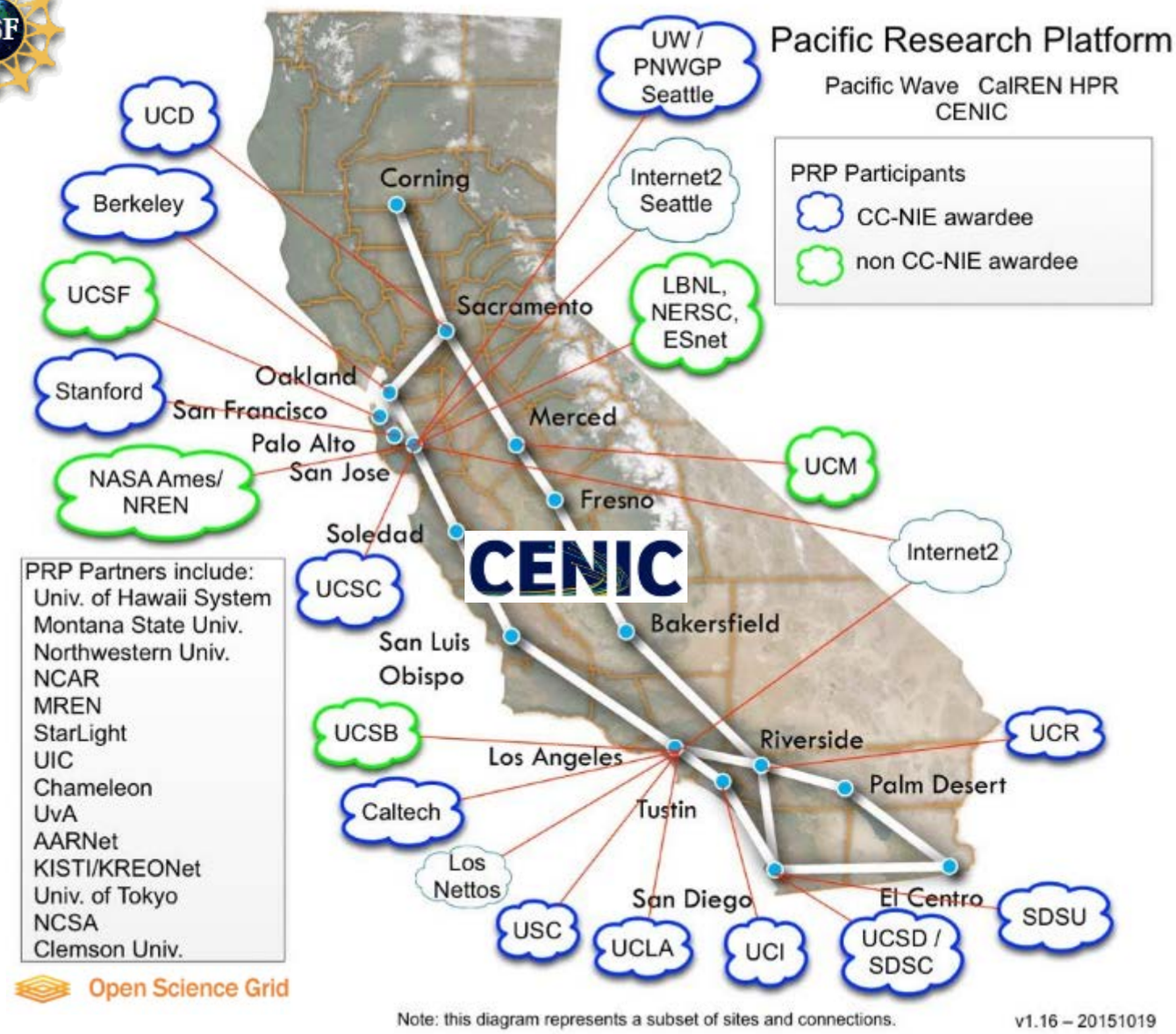
NSF Project #ACI-1541349: Larry Smarr, PI, Calit2, University of California, San Diego

PacificResearchPlatform.org

John Hess – CENIC

Thomas Hutton – SDSC/Cailit2, University of California, San Diego

Two Years In: The Pacific Research Platform v1 is Now a Working End-to-End Science-Driven DMZ-Connector



NSF CC*DNI Grant
$5M 10/2015-10/2020

PI: Larry Smarr, UC San Diego Calit2
Co-Pis:

- **Camille Crittenden, UC Berkeley CITRIS**
- **Tom DeFanti, UC San Diego Calit2**
- **Philip Papadopoulos, UC San Diego SDSC**
- **Frank Wuerthwein, UC San Diego Physics SDSC**

Science Teams:

- **Visualization and Virtual Reality**
- **Biomedical**
- **Earth Sciences**
- **Particle Physics**
- **Astronomy and Astrophysics**
- **Cryo-EM**
- **Deep Learning & Robotics**
- **High-Performance Wireless**

Source:
John Hess

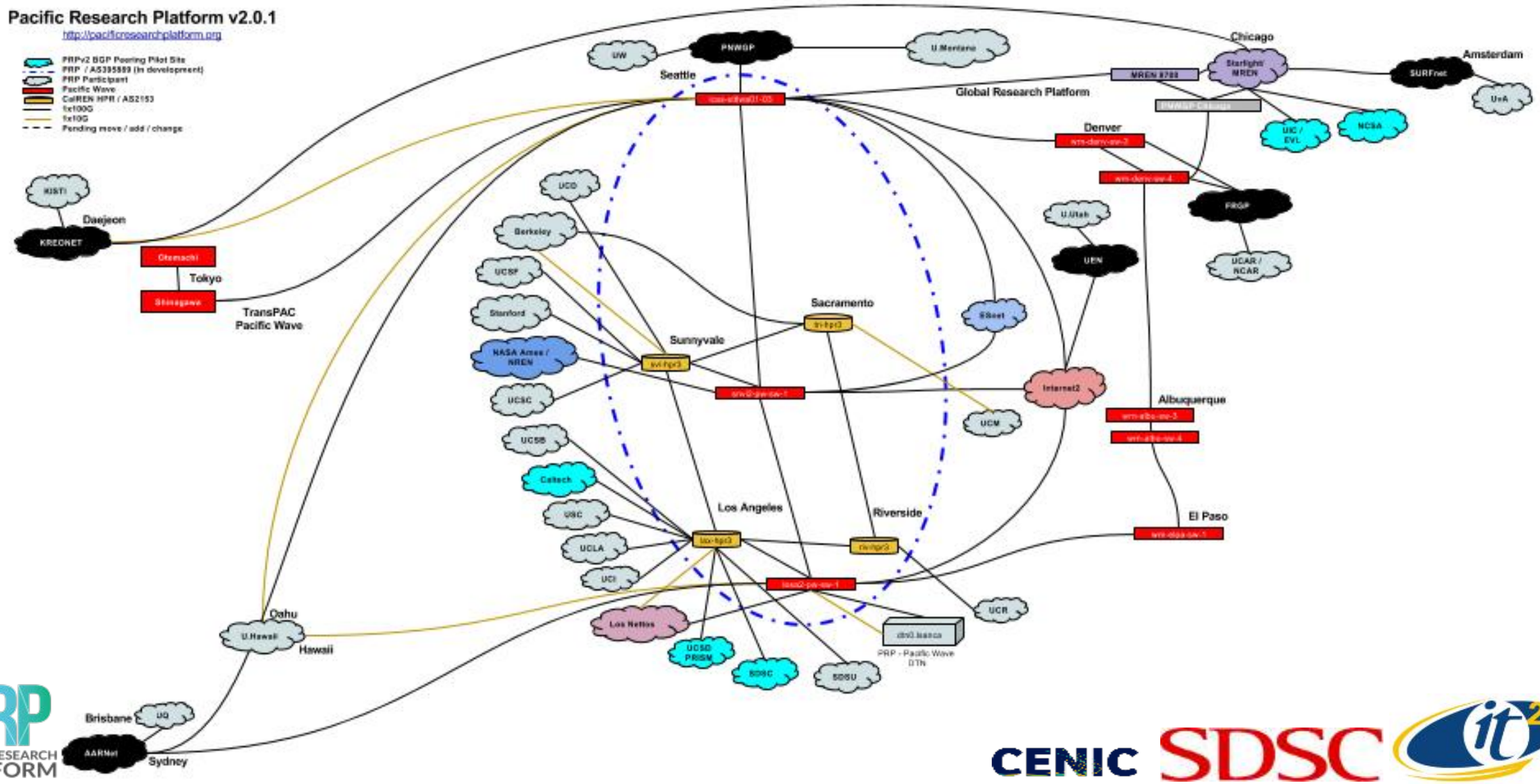
What do we envision for PRPv2

- **We have had challenges in making sure that the PRP is carrying Science DMZ to Science DMZ traffic only and is not bypassing campus firewall or IDS subsystems due to lack of tools using HPR backbone.**
  - **PRP v2 will utilize a new ASN for the PRP backbone. This ASN has been received from ARIN: AS395889**
  - **Route Servers at Pacific Wave Exchange points will be utilized for PRP resource routes**
  - **BGP Communities for tagging classes of DMZ networks**
  - **A Pilot of this BGP implementation will be created among 6 sites**
    - **Will incorporate UCSD, SDSC, Caltech, NCSA, Univ of Chicago and one yet to be determined Northern California CENIC site.**
    - **Pilot BGP Peering will be native IPv6 only. May or may not carry IPv4 as transport.**
  - **Stretch Goals**
    - **Incorporate SDN/SDX type signaling for paths or 'super-channels'**

PRP PACIFIC RESEARCH PLATFORM

CENIC SDSC i^2

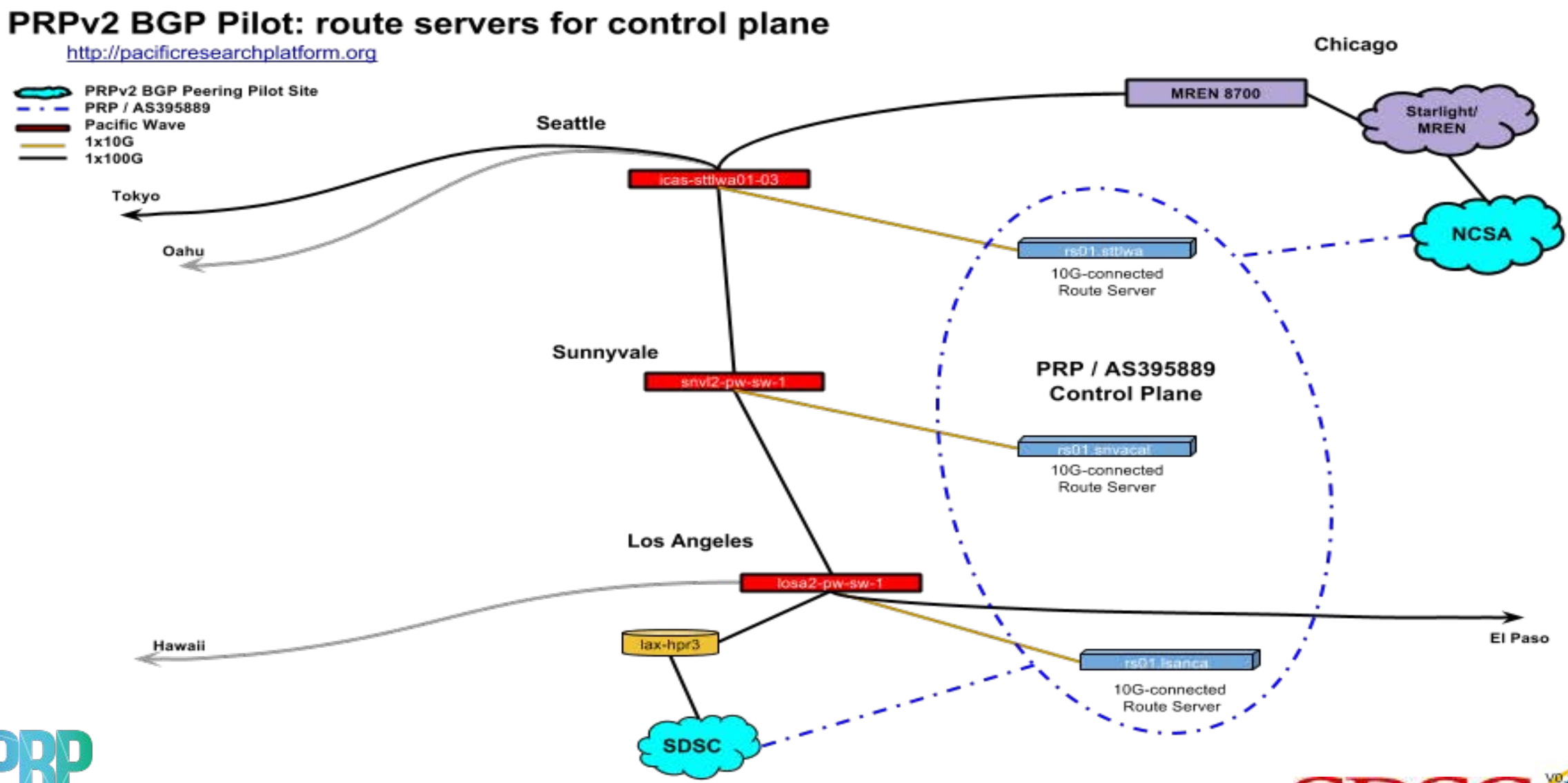
PRPv2 BGP Peering Pilot



PACIFIC RESEARCH PLATFORM

CENIC SDSC it²

PRPv2 BGP Peering Pilot



PACIFIC RESEARCH PLATFORM

CENIC

SDSC

V0.1.8
20170220

Network Evolution to IPv6: Cooperating Research Groups

Any particular science driver is comprised of scientists and resources at a subset of campuses and resource centers. We term these Science Teams with the resources and instruments they access as Cooperating Research Groups (CRGs). Resources can range from unique lab instruments, lab-owned clusters, federated resources (like OSG), to XSEDE resources. Ideally, a specific CRG should have the option of shielding access (via firewalls and authorization mechanisms) to its resources, even from a different CRG that is within an overlapping set of institutional DMZs. Achieving such isolation or controlled access (where remote systems are not allowed connection to specific PRP resources) is possible today, but has a very high network engineering and management overhead. In a real sense, members of a specific CRG trust one another, but they do not necessarily trust others (even other CRGs on the PRP). PRPv1 is generally shortened to PRP below.

Evolving to a mostly IPv6 PRPv2: A key issue with IPv4 DMZs is that endpoints on campuses have depleting pools of addresses. At institutions like UCSD, even new class C subnets (256 IPv4 addresses) are essentially unavailable. Within the current address pool, a host that is part of the PRP also could be on network segment with non-PRP hosts. This potentially complicates access control. We believe that IPv6 can provide a near-term benefit and a long-term solution. Today, most campuses have a dearth of deployed IPv6 resources. This means that inbound IPv6 traffic is more readily classified by the border router as PRP or non-PRP. The virtually unlimited space of IPv6 gives campus network administrators additional freedom in address labeling. They could, for example, assign a portion of their locally-defined IPv6 address space to indicate various PRP-subnets. Based on such a scheme, a small set of simple routing rules could reliably bifurcate PRP and campus/other DMZ traffic. We believe it is critical to define several practical schemes through which an already-overburdened campus network administrator can more easily integrate the PRP into their operations without dramatically impacting their existing IPv4 network. While IPv6 brings application challenges because some underlying software must be recoded to support v6 addressing, workhorse software like HTCondor, the perfSONAR toolkit, and OpenFlow are already IPv6-aware. The PRPv2 pilot be native IPv6 only but once in production, at least for a limited time we may need to also carry IPv4 as transport under IPv6 peering. We would classify IPv6 as successful on PRPv2 if, at the end of five years, nearly all hosts were utilizing only IPv6 on their primary network interface. PRPv2 will help accelerate adoption of IPv6.

PRP PACIFIC RESEARCH PLATFORM

CENIC SDSC it^2

ESnet 4 Public DTNs

- lbl-diskpt1.es.net, anl-diskpt1.es.net, bnl-diskpt1.es.net, cern-diskpt1.es.net
  - See: http://fasterdata.es.net/performance-testing/DTNs/
- Anonymous read-only mode
- Local RAID: 8Gbps/sec read performance
- Write access for ‘collaborators’ (ie: you!)
  - Send me your globus ID, and I can add you
- CentOS7 with FQ enabled
- Tuning settings:
  - endpoint-modify --network-use=custom --max-concurrency=32 --max-parallelism=16 --preferred-concurrency=8 --preferred-parallelism=4
  - /sbin/tc qdisc add dev eth10 root fq maxrate 8gbit
  - HTCP (not cubic)

tstat: http://tstat.polito.it/

- All ESnet and NERSC DTNs are instrumented using tstat
  - Based on libpcap capture of headers
  - requires 30-50% of 1 core (3.5Ghz Xeon) to do full line rate 10Gbps

tstat Data Collection

- 1 line per TCP socket
- SRC/DST IP and Port
- TCP stats: (each direction)
  - tcp_rexmit_bytes , tcp_rexmit_pkts, tcp_rtt_avg, tcp_rtt_min, tcp_rtt_max, tcp_rtt_std, tcp_pkts_rto, tcp_pkts_fs, tcp_pkts_reor, tcp_pkts_dup , tcp_pkts_unk, tcp_pkts_fc, tcp_pkts_unrto , tcp_pkts_unfs, tcp_cwin_min, tcp_cwin_max, tcp_out_seq_pkts, tcp_window_scale, tcp_mss , tcp_max_seg_size, tcp_min_seg_size , tcp_win_max, tcp_win_min, tcp_initial_cwin
- Think of this as Netflow ++

tstat analysis

- **Goal is to generate a report:**
  - List of the largest N flows where throughput appears to be limited by CWIN (ie: host tuning needed)
  - List of the largest N flows where throughput appears to be impacted by packet loss (‘rexmit pkts’)
  - List of the largest N flows where throughput appears to be impacted by ‘reordering’
  - List of the largest N flows where throughput appears to be impacted by ‘net dup’
  - Etc.