

---

# Network Research Efforts

---

Phil DeMar  
March 7, 2012



---

# E-Center



**Office of  
Science**

U.S. DEPARTMENT OF ENERGY



---

# Project Overview

- Network Path Weather Map Service
  - User-friendly tools to help isolate network path problems
  - Help set network performance expectations
- Social Portal for the DOE User Community
  - Place for network users to share issues, ideas, concerns
- Scoped to incorporate the DOE (SC) Enterprise
- Project Partners:
  - ESnet (PerfSONAR infrastructure & development)
  - Bucknell (Forecasting)
  - Ohio Supercomputing Center (Anomaly Detection)



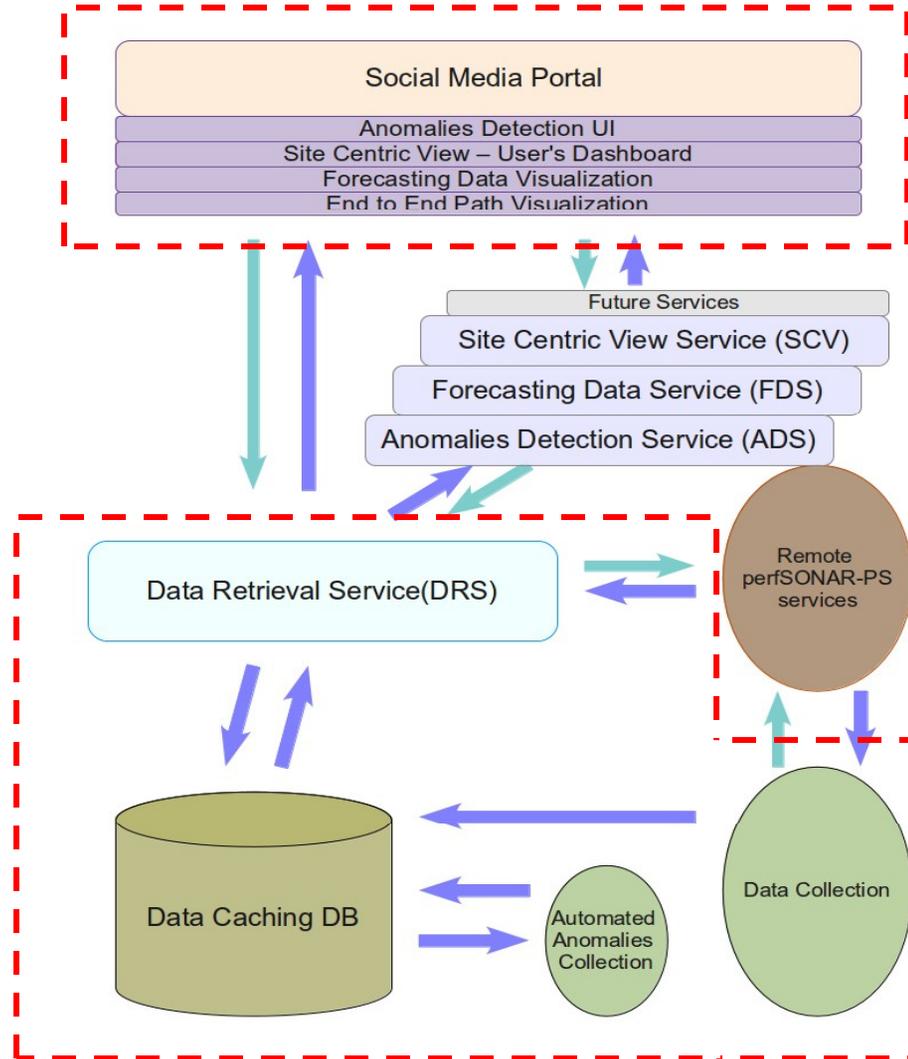
---

# Project Details

- Three-year DOE ASCR-funded project
  - Scheduled to end in September (2012)
- Maxim Grigoriev is project manager
  - Frontend contractor left last month
  - New contractor likely to start in 2 weeks
- FNAL budget is \$350k/yr
  - On target to fund Max (100%) for FY12 & 6 months of contractor
  - Project partners funded at much lower (\$150k/yr) level



# E-Center's Major Components



**Front End**  
**User Interface**  
**Data Visualization**

**(Added Services)**

**Back End**  
**Data Collection (PerfSONAR)**  
**Data Caching**  
**& Rationalization**



# Data Collection

- **perfSONAR-PS** Metadata collected and cached – nightly
  - From ESnet hLS – home Lookup Services
  - From hLSs at DOE Labs
- Data collected and cached – nightly for the past 24 hours
- From all **DOE labs** and **ESnet** measurement mesh:
  - OWAMP (delay, packet loss)
  - BWCTL (throughput)
  - PingER (RTT, packet loss)
- From ESnet centralized services:
  - Interface Utilization, Drops, Errors
  - ESnet Topology
  - Dynamic Circuits Topology and Utilization
  - Traceroutes between ESnet PoPs and DOE lab's



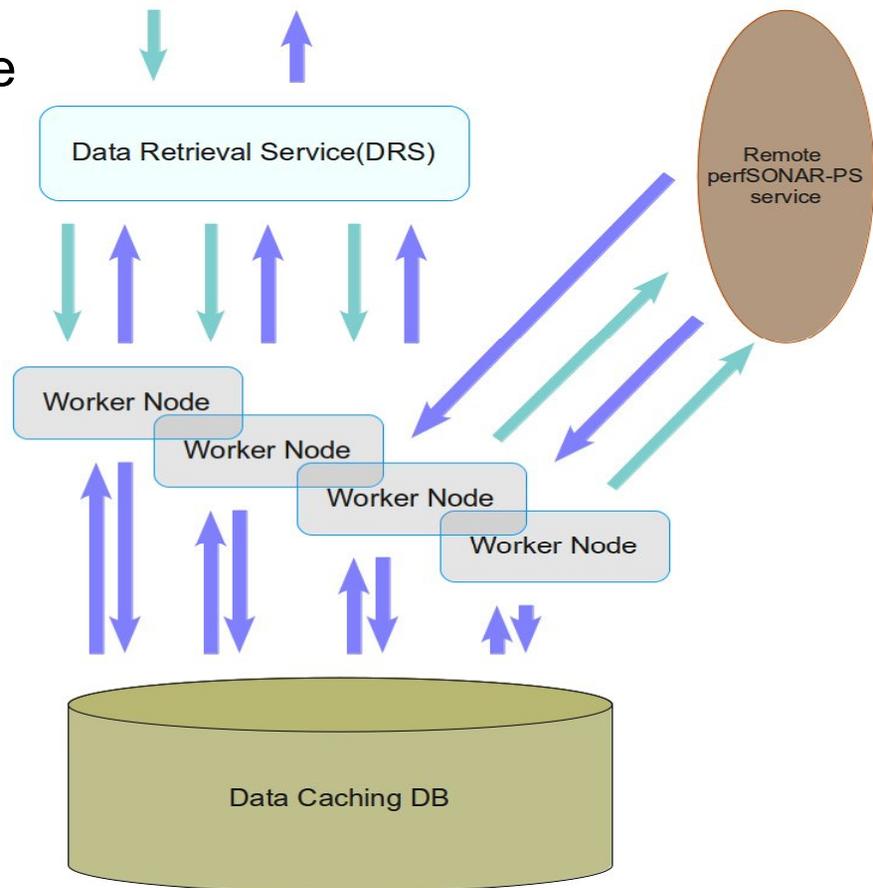
# Data Collection: Just Numbers

- Monthly cached data – entries range:
  - OWAMP 60-100M+
  - SNMP 35-50M
  - PingER 10-20M
  - Traceroute 0.6-1.2M
  - BWCTL 60K-120K
- Collecting data from ~300 service endpoints, about 25000 unique measurements (metadata entries).
- Nightly automated anomalies detection for BWCTL and OWAMP data – about 9K metadata entries weekly.



# DRS: Data Retrieval Service

- Scalable, extendable web service
- ■ Think of it as PerfSONAR middleware...
- Returns cached data or data from remote pS-PS services
- End to End tomography via SNMP & Traceroute
- Parameters:
  - Time period
  - Source/Destination IPs
  - Source/Destination HUBs
  - Data type
  - User's provided Traceroute
  - Resolution - number of aggregated data points

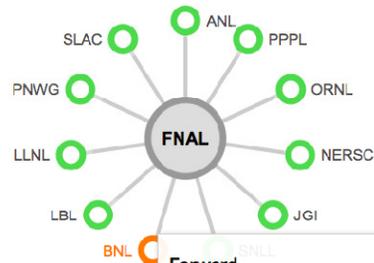


# Frontend: Dynamic, User-Configurable Dashboard

## Dashboard

### FNAL status

2011-09-08 04:00:00  
to  
2011-09-08 10:00:00



### Tools

[Weathermap](#)      [Network health](#)  
[Create issue](#)      [Create wiki page](#)

### Hot conversations

The most active public conversations in the past 2 weeks

#### [Hello, conference attendees!](#)

by [eads](#) | 6 recent comments  
Last comment: 2 weeks 5 days ago

### Most popular content

The most popular content in the past month

#### [FNAL to BNL, 2011-06-06 05:00:00 - 2011-06-07 05:00:00 \(UTC\)](#)

by [maximgrp](#) | Issue  
Last updated: 1:34pm 9-07-2011

#### [Hello, conference attendees!](#)

by [eads](#) | Issue  
Last updated: 4:20pm 7-14-2011

### Site activity

Activity on the site and in your groups



Issue [FNAL to BNL, 2011-06-06 05:00:00 - 2011-06-07 05:00:00 \(UTC\)](#) updated by [maximgrp](#)

[maximgrp](#) wrote: *There is a higher than normal one way*



[admin](#) commented on [Hello, conference attendees!](#) 3 weeks, 1 day ago

[admin](#) wrote: *Replied from Outlook web express.*



Wiki page [Data Retrieval Service Specification](#) updated by [admin](#) Jul 14, 2011

Log message: *Current version is 3.3*

#### Forward

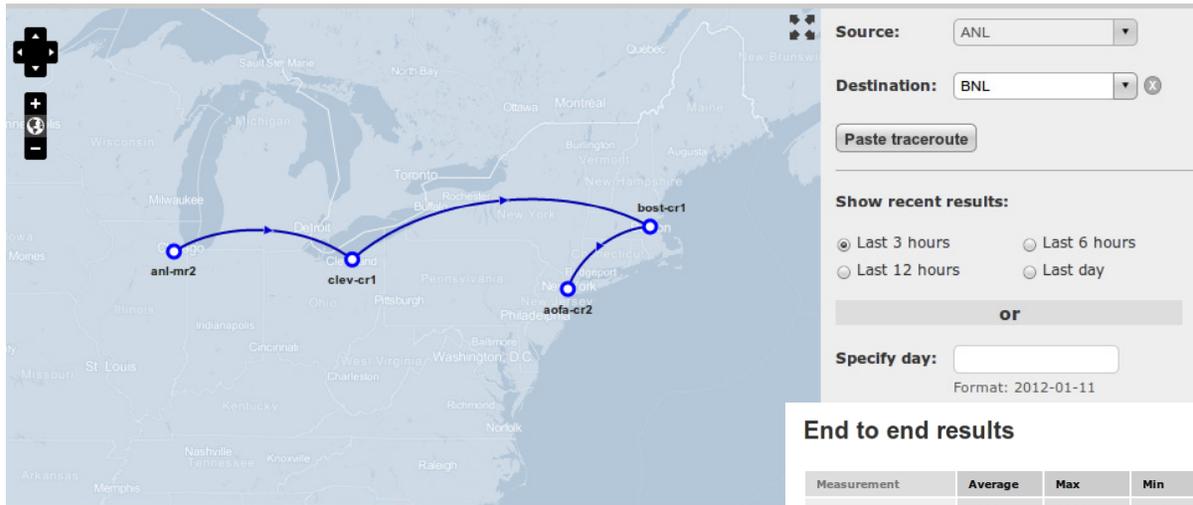
Measurement	Value	Interface
Utilization	89.22	134.55.221.138
Errors	0.00	NA
Drops	0.00	NA

#### Reverse

Measurement	Value	Interface
Utilization	85.64	134.55.41.121
Errors	0.00	NA
Drops	0.00	NA



# Front-end: E2E UI & data visualization



## ANL to BNL

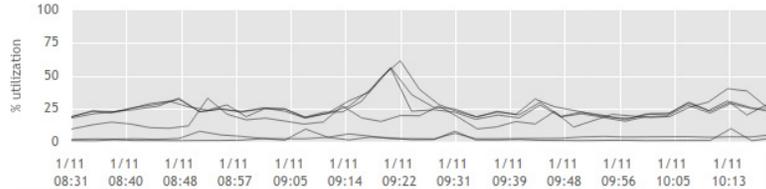
Jan 11, 2012 08:30 to Jan 11, 2012 11:30

Type: **SNMP** [Get forecast](#)

Traceroute (logical) ■ Forward ■ Reverse [Click nodes to see detail](#)



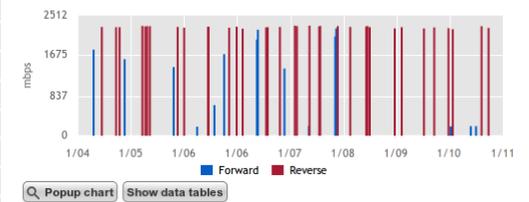
### Per hop utilization



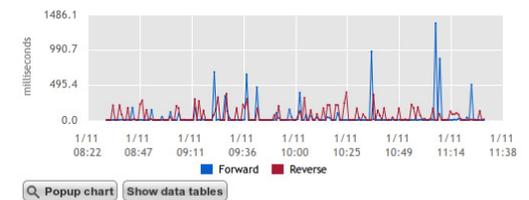
### End to end results

Measurement	Average	Max	Min
Forward Throughput	1202mbps	2228mbps	186mbps
Reverse Throughput	2257mbps	2284mbps	2213mbps
Forward Duplicates	0	0	0
Reverse Duplicates	0	0	0
Forward Loss	0.0	0.0	0.0
Reverse Loss	0.0	0.0	0.0
Forward Max delay	58ms	1351ms	13ms
Reverse Max delay	67ms	396ms	13ms
Forward Min delay	12ms	13ms	12ms
Reverse Min delay	13ms	13ms	12ms

### Throughput Jan 4, 2012 10:38 - Jan 11, 2012 10:19



### Max delay Jan 11, 2012 08:31 - Jan 11, 2012 11:29



# Year 3 Plans & Deliverables

## ■ Year 3 Objectives:

- ❑ Hardening E-Center data retrieval system
- ❑ Circuit-based network path monitoring - *completed*
- ❑ On-demand measurement test capabilities:
  - Traceroute
  - BWCTL
- ❑ Alerts and updates notifications framework - *completed*
- ❑ Work with DOE labs and roll out wide-scale deployment of the perfSONAR-PS services under the E-Center management.
- ❑ Outreach to seed use of social portal by user groups:
  - IPv6
  - Video/collaborative tools

## ■ Overall Project Deliverable:

- ❑ Functional E-Center service...



---

# Future Directions/Ideas

- Standalone DRS
  - Deployable at end-sites & other domains
  - Facilitate domain data protection and anonymization
  - Greater flexibility in supporting roles-based end-to-end data access
- Integration or/and collaboration with myESnet portal
  - Leveraging developed components and technologies
  - Efficient utilization of resources
  - Single Point of network measurement information for ESnet users



---

# Future Related Funding Opportunities

- Standalone DRS with full AAA stack & federation capability
  - A missing element in PerfSONAR monitoring/measurement development
  - A potential generic network infrastructure service for any application needing network data
- Integration work into myESnet portal platform
  - Would depend on scope of integration effort
  - Possibly propose to provide DRS-based development environment



---

# Project Information

- E-Center: <https://ecenter.fnal.gov>
- E-Center Project Site, including documents & software repository:

<https://cdcv.s.fnal.gov/redmine/projects/ecenter>



---

# End Site Control Plane System

(ESCPS)



# Project Overview

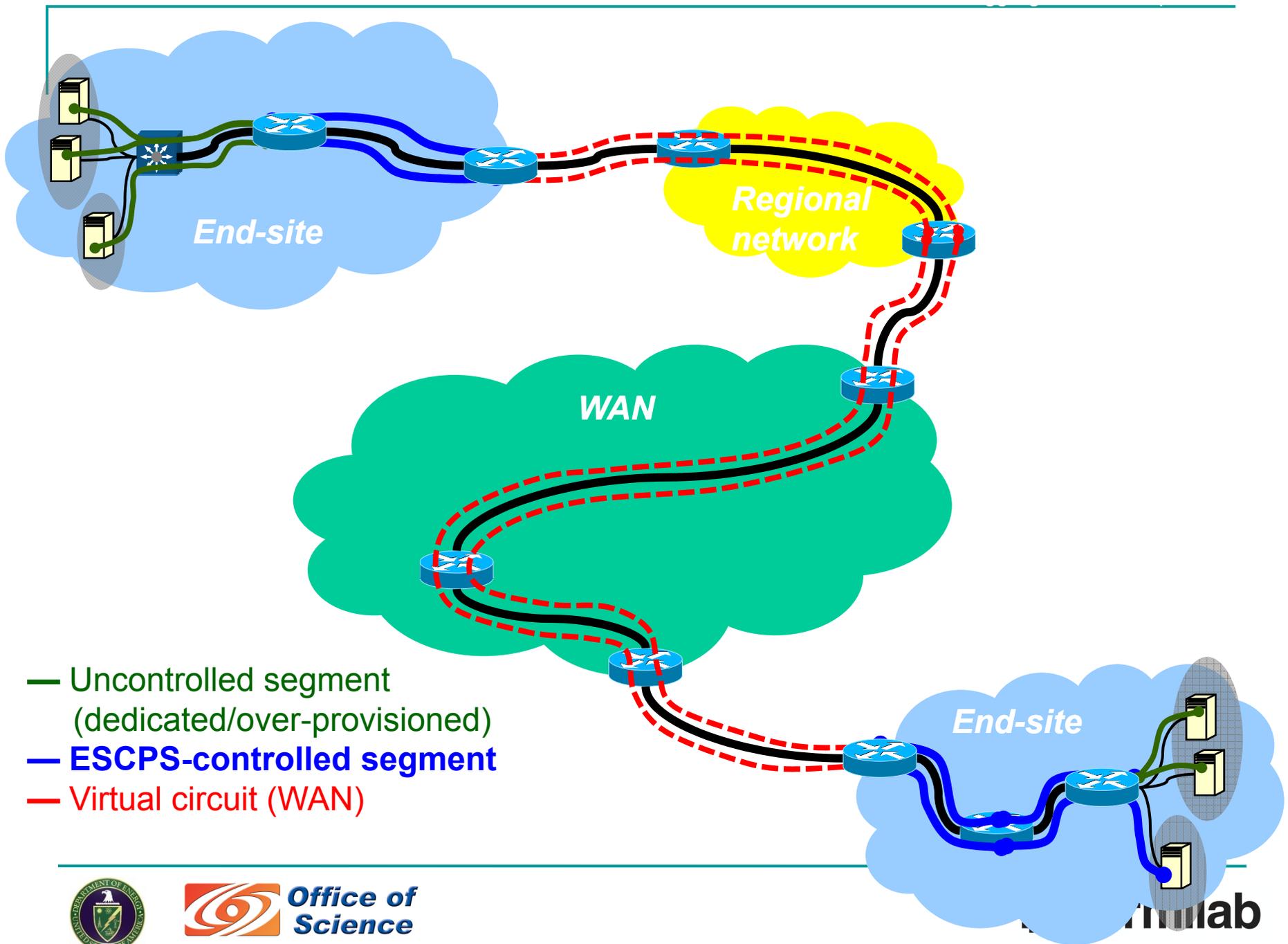
- End site control component of emerging integrated dynamic circuit services
  - Extension of ESnet OSCARs circuit service into end sites
  - Vision is the end site element of federated control plane capability
  - Provide end-to-end circuit support capabilities
- Expected to build on earlier work on LambdaStation (FNAL)
  - And similar legacy site-based circuit services (TeraPaths, Phoebus)
- End-sites have end-to-end perspective:
  - End-to-end (host-to-host) paths
  - Circuit management
  - Distributed reservations
- Project Partners:
  - BNL (TeraPaths...)
  - Delaware (Phoebus...)



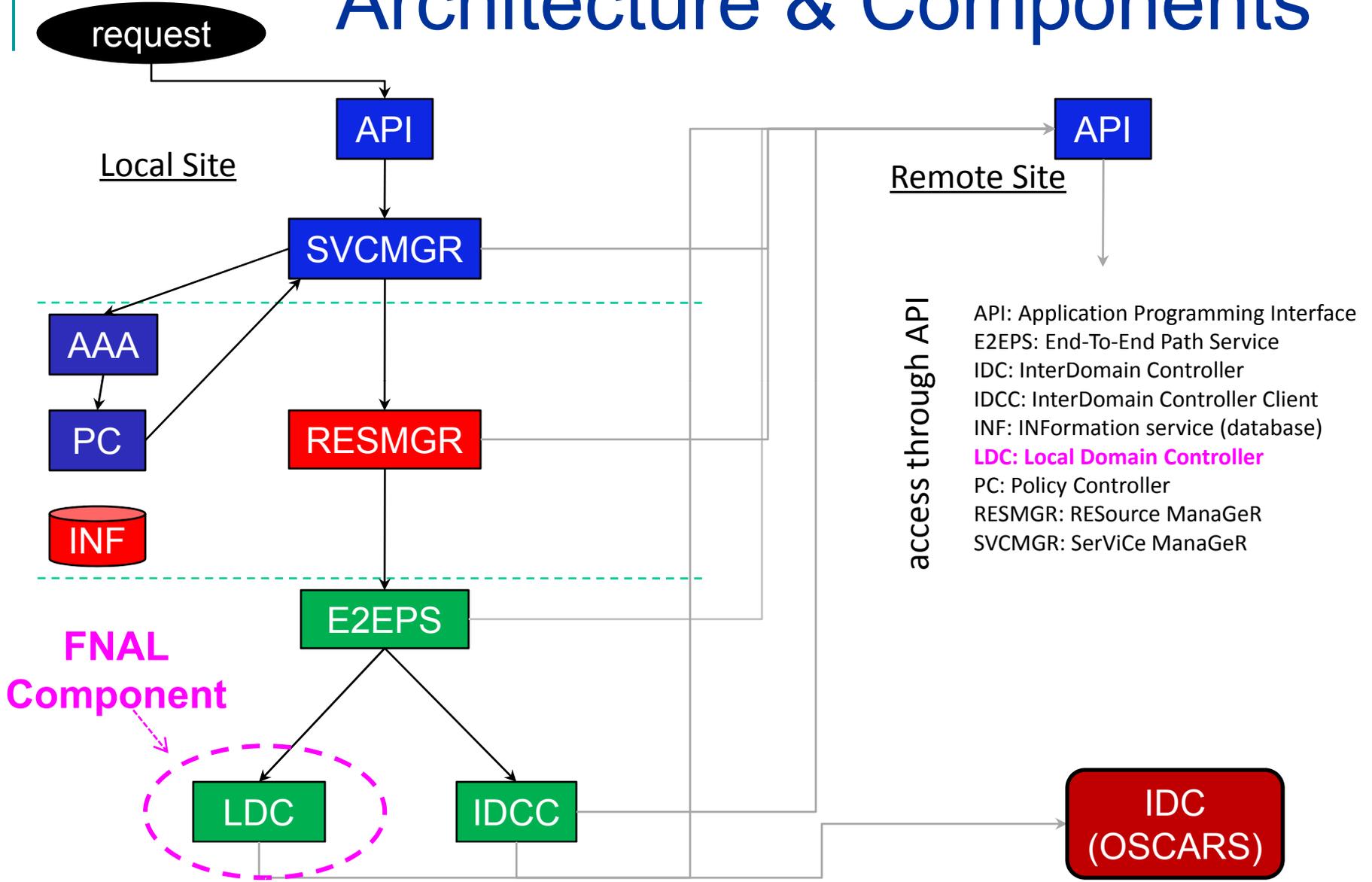
# Project Details

- Three-year DOE ASCR-funded project
  - Scheduled to end in September (2012)
- Andrey Bobyshev has led FNAL's design effort
  - Contractor left last summer
  - Trying to utilize in-house resources for remaining scripting
- FNAL budget is \$180k/yr
  - On target to fund 6 months of developer (internal or external)
  - Andrey's design effort basically done
    - ~10% effort in oversight of developer & coordination with BNL
- Project partners were equally funded
  - Delaware not funded for year 3
    - this is causing overall project effort problems





# Architecture & Components



---

# Component Assignments

- BNL:
  - ❑ Resource manager
  - ❑ Inter-domain controller (IDC) interface
  - ❑ Coordination of integration effort
- FNAL:
  - ❑ Local domain controller (LDC)
  - ❑ Initial service manager (SVCMGR) design
- U-Del:
  - ❑ Local network monitoring
  - ❑ Service manager (SVCMGR) adaptation for XSP



---

# FNAL Component of ESCPS



**Office of  
Science**

U.S. DEPARTMENT OF ENERGY



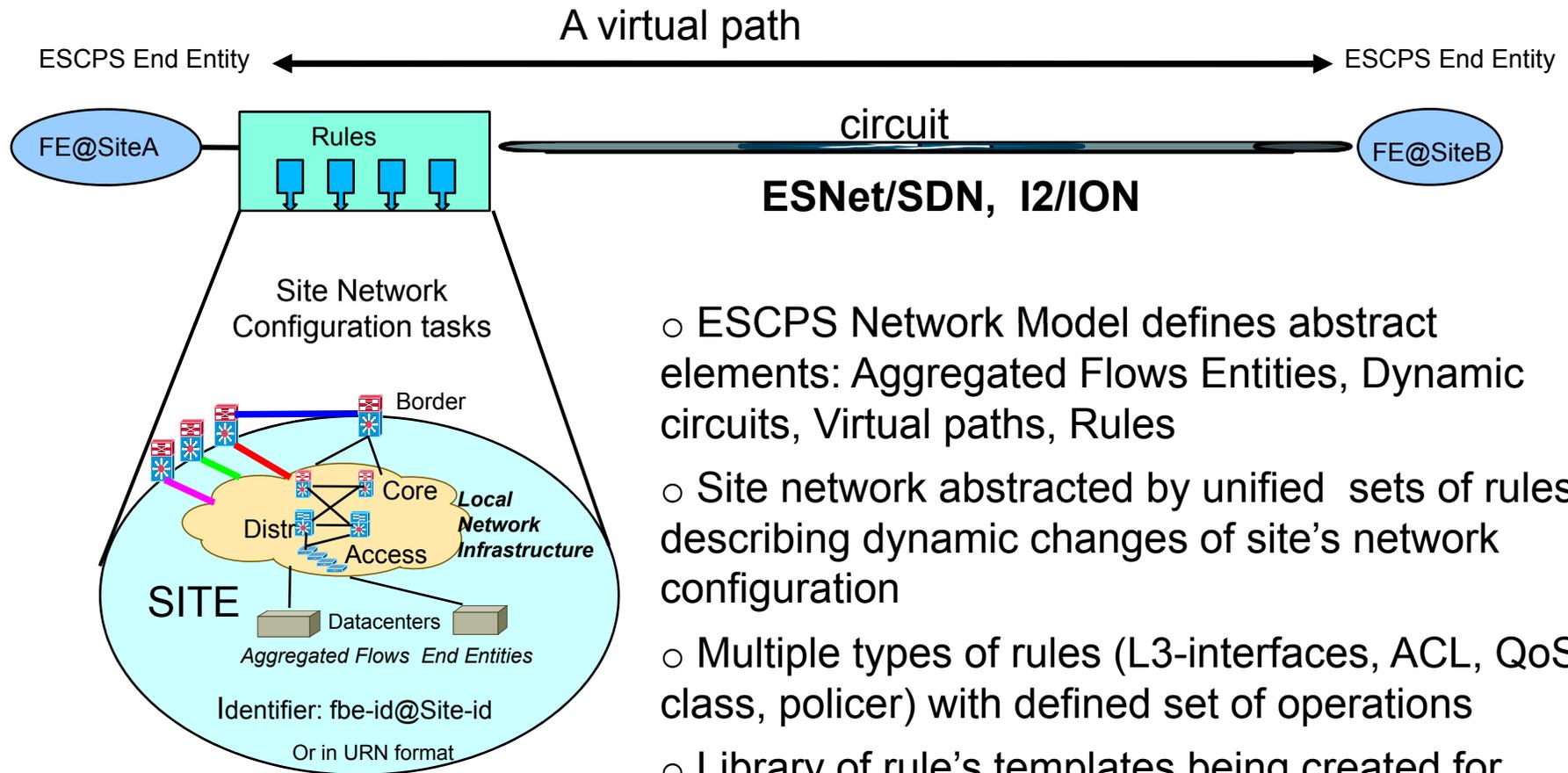
---

# Local Domain Controller (LDC)

- Configures LAN component of end-to-end path on-demand
- Technology agnostic, create a virtual path within LAN:
  - As specified by site
  - Based on local network technology deployed
- Describes local network path in terms of a finite number of configuration rules, unified interface
- Interacts with RESMGR to obtain information on flows in submitted tickets
  - Signals RESMGR when job is done or in case of failures
  - RESTFul web interface to access LDC services



# Local Path Network Model



- ESCPS Network Model defines abstract elements: Aggregated Flows Entities, Dynamic circuits, Virtual paths, Rules
- Site network abstracted by unified sets of rules describing dynamic changes of site's network configuration
- Multiple types of rules (L3-interfaces, ACL, QoS class, policer) with defined set of operations
- Library of rule's templates being created for selected network technology used at different sites



---

# LDC Service Manager

- Workflow manager for LDC:
  - Provides information about available services, sites, virtual paths
  - Functionally a front-end for LDC services
- Accessible via RESTful Web Services
- Interfaces with other ESCPS components via a simple event-based mechanism
  - Service requests processed based on the defined workflow.
  - Each state in workflow can be configured with a checklist of awaiting events from external processes
- Java package in beta stage
  - Installation package includes all components needed in Linux environment



escps@fnplug.fnal.gov

escps@fnplug.fnal.gov

http://fnplug:9000/resources.xhtml

Most Visited Getting Started Latest Headlines N Naqios N Manmon ActiveMQ Console Cacti https://fnab3.fn...

Logged in as: **andrey** Logout

# ESCPS@ FNPLUG.FNAL.GOV

END SITE CONTROL PLANE SUBSYSTEM

[View Tickets](#) [Create New Ticket](#) [Administration](#) [Monitor](#)

Status:  (ACCEPTED, ACTIVATING, ACTIVE, BOOKED)

Start Time:

End Time:

Ticket #:

Requestor User Name:

### Active Tickets

Ticket Id	Reservation Id	Ticket Status	Reservation Status	Remote Ticket Id	Remote Ticket Status	Local Entity	Remote Entity	Application Class	Start Time	End Time	Requestor	Chart	Cancel
667	688	Started	Active	16		USCMS-T1	kazim	Streaming	06/23/11 16:41:41	11/09/13 02:21:41	test	Display	Cancel
668	689	Started	Active	67		USCMS-T1	kazim	Real-Time	06/23/11 16:42:00	12/12/13 10:22:00	test	Display	Cancel
669	690	Started	Active	79		USCMS-T1	kazim	Bulk Scavenger	06/23/11 16:42:05	04/27/12 17:22:05	test	Display	Cancel
670	690	Started	Active	23		USCMS-T1	kazim	Bulk Scavenger	06/23/11 16:42:13	03/04/13 02:22:13	test	Display	Cancel
671	691	Started	Active	98		USCMS-T1	kazim	Data Movement	06/23/11 16:42:13	02/02/13 22:22:13	test	Display	Cancel
672	692	Started	Active	14		USCMS-T1	kazim	Default	06/23/11 16:42:16	10/14/14 00:42:16	test	Display	Cancel
673	692	Started	Active	163		USCMS-T1	kazim	Default	06/23/11 16:42:23	04/24/12 06:02:23	test	Display	Cancel
674	692	Started	Active	42		USCMS-T1	kazim	Default	06/23/11 16:42:26	04/03/14 23:22:26	test	Display	Cancel
675	692	Started	Active	315		USCMS-T1	kazim	Transactional	06/23/11 16:42:51	09/27/11 12:42:51	test	Display	Cancel
676	689	Started	Active	223		USCMS-T1	kazim	Real-Time	06/23/11 16:43:15	09/23/11 08:43:15	test	Display	Cancel

118 tickets found, displaying 10 tickets, from 1 to 10. Page 1 / 12

# LDC Service Manager



# LDC Status & Year 3 Deliverables

- Current Status
  - Design is completed (XML schemes, Rule templates, Network model examples in XML)
  - LDC workflow manager in prototype stage
  - Finished LDC rules for Catalyst 6509/IOS
  - Developing LDC rules for NX-OS, Cisco Nexus 7000/5000
  - LDC device configuration development being done in Perl
- FNAL project deliverables:
  - Complete LDC, with device templates for Cisco platforms
  - Work w/ BNL to integrate LDC component into RESMGR platform
- Investigate OpenFlow support for LDC
  - Not a project deliverable...



---

# Future Related Funding Opportunities

- ESCPS/LDC adaptation for OpenFlow
- Integration of ESCPS into OSCARs platform



---

# Project Information:

## ESCPS Documents:

<https://plone3.fnal.gov/P0/ESCPS>

## ESCPS Software TRAC repository:

<https://damsl.cis.udel.edu/escps>



---

## Other Network Research Efforts

- 1) GPU-Based network monitoring & performance analysis
- 2) Host-optimizations for network I/O intensive environments

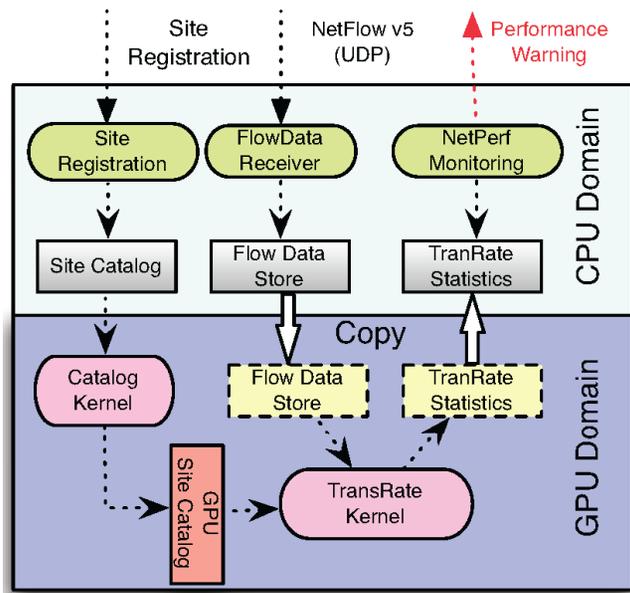


# G-NetMon

- A GPU-accelerated network performance monitoring system to support large-scale scientific collaborations
- Original prototype - analyzed network flow data to calculate transfer rates between Fermilab and collaboration sites.
  - fast analysis of bulk data movement
  - rapidly detect sub-optimal bulk data movements.
- Demonstrated that GPU can process millions of flow records in seconds
- Papers:
  - Wenji Wu, Phil DeMar, Donald J. Holmgren, Amitoj Singh, Ruth Pordes: G-NetMon: A GPU-accelerated network performance monitoring system for large scale scientific collaborations. LCN 2011: 195-198.
  - Wenji Wu, Phil DeMar, Donald Holmgren, Amitoj Singh, G-NetMon: A GPU-accelerated network performance monitoring system, The proceedings of SAAHPC'11 symposium, Knoxville, TN, 2011, pages: 76-79.



# G-NetMon Design



**G-NetMon Architecture**

The G-NetMon architecture consists of a few parts that are executed on either the host (CPU) or GPU. Based on the CUDA design principle, the parts that exhibit little or no data parallelism are implemented as sequential CPU threads; the parts that exhibit a rich amount of data parallelism are implemented as GPU kernels.

## CPU Domain

- Site Registration Thread
- FlowData Receiver Thread
- NetPerf Monitoring Thread

## GPU Domain

- Catalog Kernel
- TransRate Kernel



---

# G-Netmon Future Directions

- Prototyping G-Debugger:
  - A GPU-accelerated network traffic monitoring & analysis system to support large-scale scientific collaborations
  - Monitors and analyzes network traffic at packet level
  - Could quickly identify root cause of a performance problem
  - Single system that monitors and analyzes network traffic @20Gb/s
- Targeted at 40GE/100GE networking environment
- Envisioned as a generic high performance packet analysis service:
  - Adaptable for other 40GE/100GE network applications



# Multicore Communication Architecture (MCA)

- Aim - design advanced network protocol stacks and OS mechanisms to perform efficient multicore network I/O
  - reduce data access delay in distributed computing environments
- Proposes new OS kernel design separating communication from computation
  - maximize parallelism between communication and computation for mitigating data access delay effects.
- Studies new host protocol stack mechanisms to streamline and speed up protocol processing
- (Related) Papers:

Wenji Wu, Phil DeMar, Matt Crawford: Sorting Reordered Packets with Interrupt Coalescing. *Computer Networks* 53(15): 2646-2662 (2009)



---

# MCA Future Directions

- Proposal to develop prototype of SL optimized for HP network I/O environments:
  - Targeted at 40GE-connected host environment
  - Align with other network research-funded projects
- May pursue X-Stack funding
  - This would likely require an xLCF partner...



# A-Transport-Friendly NIC (A-TFN)

- New NIC mechanism for multi-core systems
- Remedies inefficiencies with existing NIC technologies
  - Receive Side Scaling RSS
  - Intel Flow Director
- Steers incoming packets to same core as their application
  - ensures in-order packet delivery (increasing problem at 40/100GE)
  - supports efficient, scalable network I/O operations for multi-core systems
- Papers:

Wenji Wu, Phil DeMar, Matt Crawford: Why Can Some Advanced Ethernet NICs Cause Packet Reordering? IEEE Communications Letters 15(2): 253-255 (2011)

Wenji Wu, Phil DeMar, Matt Crawford: A Transport-Friendly NIC for Multicore Systems. IEEE Transactions on Parallel and Distributed Systems, 2012



---

# A-TFN Future Directions

- Unlikely to submit as a stand-alone proposal:
  - Doesn't seem to be much interest within ASCR Network R&D in NIC technologies
  - May also pursue X-Stack funding here
    - Again, would likely require an xLCF partner...
- A-TFN does complement MCA's target optimizations
  - Could be implemented in software
  - May include it within an MCA proposal



---

# Collaborative Efforts w/ SC/SCF

- CortexNet:

- Joint white paper with DOCS on network-awareness service for data locator service
- Target data locator service = XrootD
- Would provide intelligent data source selection based on network path capabilities & conditions
  - Capitalize on E-Center DRS development
- Overall objective = framework for proposal response to solicitations
- Not clear what solicitation venue would be appropriate



---

# 100GE Status



**Office of  
Science**

U.S. DEPARTMENT OF ENERGY



# ESnet-provided 100GE Connectivity to FNAL

- 100GE-capable ESnet ChiExpress (MAN-II) hub in burn-in on FCC2E now...
  - Likely turn-over to ESnet in April
  - ESnet could provide us a 100GE interface ~July
- ESnet to migrate existing services to new infrastructure by Thanksgiving
  - Not clear how this will impact ChiMAN (MAN-I) waves
  - Likely several channels will remain for redundancy
- Need 100GE router/interface to connect to ESnet @100GE
  - Current state of 100GE router/switch market being investigated...
  - Also need to analyze existing LAN/WAN architecture for how best to accommodate 100GE technology



---

# ESnet ANI Test Bed

- ESnet 100GE test bed funding ends in August
  - Parts of existing test bed are being migrated to production network
  - ESnet will be submitting unsolicited proposal to extend test bed
    - But, its not clear that additional funding will be provided
- FNAL could join 100GE test bed with a local 100GE router
  - Expect B/W split between production & test bed
    - Already using 7 x 10GE for production
  - Caveat emptor wrt disposition of ANI test bed



# Local FNAL 100GE Test Bed Possibilities

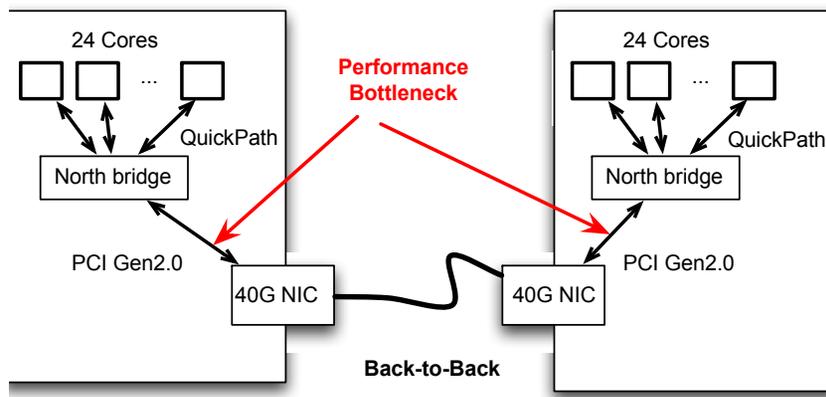
- Local networking resources available for ANI test bed participation:
  - Cisco 6509 router (n x 10GE...)
  - Two HP systems w 40GE NICs
    - But minimal disk & no RAID
    - A motley collection of 5-6 year old 1GE and 10GE-connected systems
- High Throughput Data Program (HTDP)
  - Proposal (DOCS) to develop local test bed infrastructure for ANI
  - Inter-Division effort with networking/monitoring support from NVS/NR
  - Objective to test middleware services at 40/100GE level



# 40GE Host Testing & Evaluation

- Currently evaluating 40GE host/NIC technologies
  - Software
    - TCP/IP stack, OS memory subsystem, middleware, application
  - Hardware
    - Mellanox NICs, PCI-E Gen-2 system bus, 12 cores

**Best performance = ~19Gbps**



Factors impacting performance:

System bus:

PCI-E Gen2 (26Gb/s) system bus

PCI-E Gen3 system bus not available yet

40G NIC:

40G NIC technology is not mature working with Mellanox to try to improve the performance



---

# Additional Slides



**Office of  
Science**

U.S. DEPARTMENT OF ENERGY



---

# E-Center Extra Slides



**Office of  
Science**

U.S. DEPARTMENT OF ENERGY



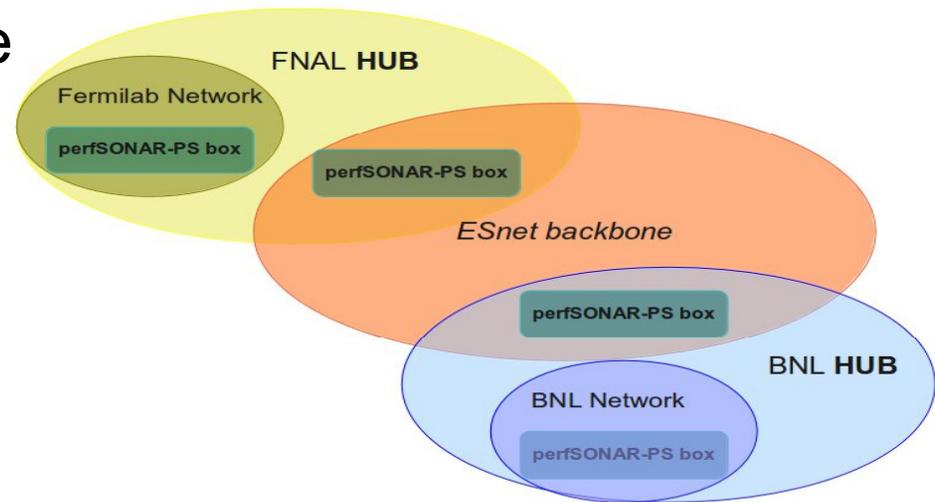
# Technical Details

- Choice of the **REST** Data services – simple, flexible
- Data cache - **MySQL DB** engine – scalable, supports replication and clustering for the future extension, ***data sharding*** – slicing each data set in pieces for fast retrieval – we use a ***timestamp*** for monthly tables
- **Gearman** distributed Job submission manager for all backend tasks – to provide desired level of scalability
- **Modern Perl** for the backend and services – stable, 20+ yr old language, great variety of APIs, mature web development framework, compatibility with perfSONAR-PS API



# DRS: What is the HUB ?

- Logical aggregation of the End-to-End pS-PS monitoring endpoints - End Site located with ESnet edge located
- Allows one to see all available data for WAN network path
- Every pS-PS service and as result every measurement is mapped to specific HUB upon discovery



# Front-end

- Powered by Drupal Content Management System
  - Popular and proven: Great community, stable, widely-known. Strong support for social networking features (issue tracking, blogs, knowledge base, tag cloud)
  - Widely adopted by government: Challenge.gov, Whitehouse.gov, Energy.gov, Recovery.gov
- Key social features
  - Data conversations: Issue & data notification system allows conversation & collaboration across DOE networking community.
  - Knowledge base: A powerful WIKI with an easy editor for generating and sharing documentation.
  - Groups: Lightweight system for sharing conversations & documentations with a task force or working group
  - Activity stream: A birds' eye view of important conversations, anomalies, and knowledge base articles customized for each user



# Authentication & User's Registration

- Make signup/login easy AND safe
- Support for Single Sign-On AA systems deployed at DOE: **OpenID**, **Shibboleth** (experimental)
- Supports common OpenID providers (Google, Facebook, etc.) as well
- Fallback to domain-based white-lists and manual user management

## User account

Click your account provider:



Username: \*

Enter your E-Center username.

Password: \*

Enter the password that accompanies your username.

Log in



---

# Front-end: Dashboard

Dynamic, per-user dashboard:

- ❑ encourages users to interact with site and each other
- ❑ user-selectable site-centric view and other data tools to give user instant snapshot of network status
- ❑ site-wide activities using familiar "activity stream" interface
- ❑ Activity stream includes detected anomalies, new content, comments, and more, based on user's group membership



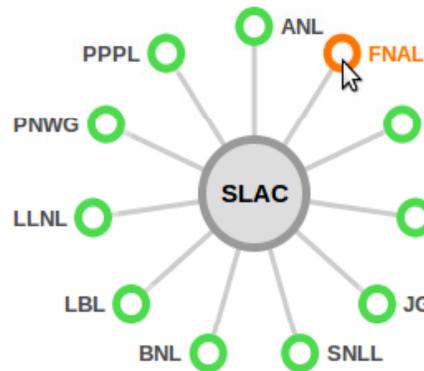
# Front-end: Site-centric view

- User-customizable view of connections to other DOE sites.
- 24-hour snapshot to find where problems exist
- Cached hourly for speedy display on dashboard

## E-Center dashboard

### SLAC status:

2011-07-05 04:00:00 -  
2011-07-05 16:00:00



### Tools

#### Forward

Measurement	Value	Interface
Utilization	87.12	134.55.219.122

Errors 0.00 NA

Drops 0.00 NA

#### Reverse

Measurement	Value	Interface
Utilization	55.51	134.55.209.45

Errors 0.00 NA

Drops 0.00 NA

8 recent comments

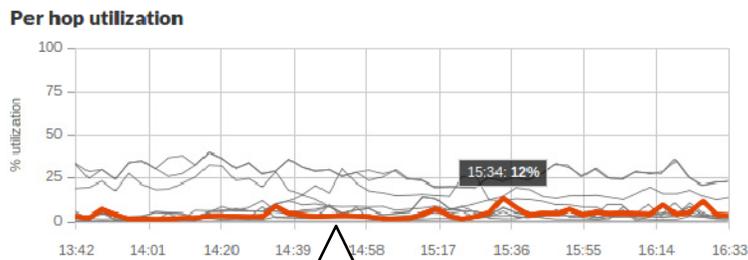
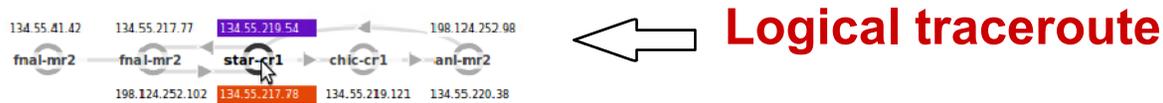
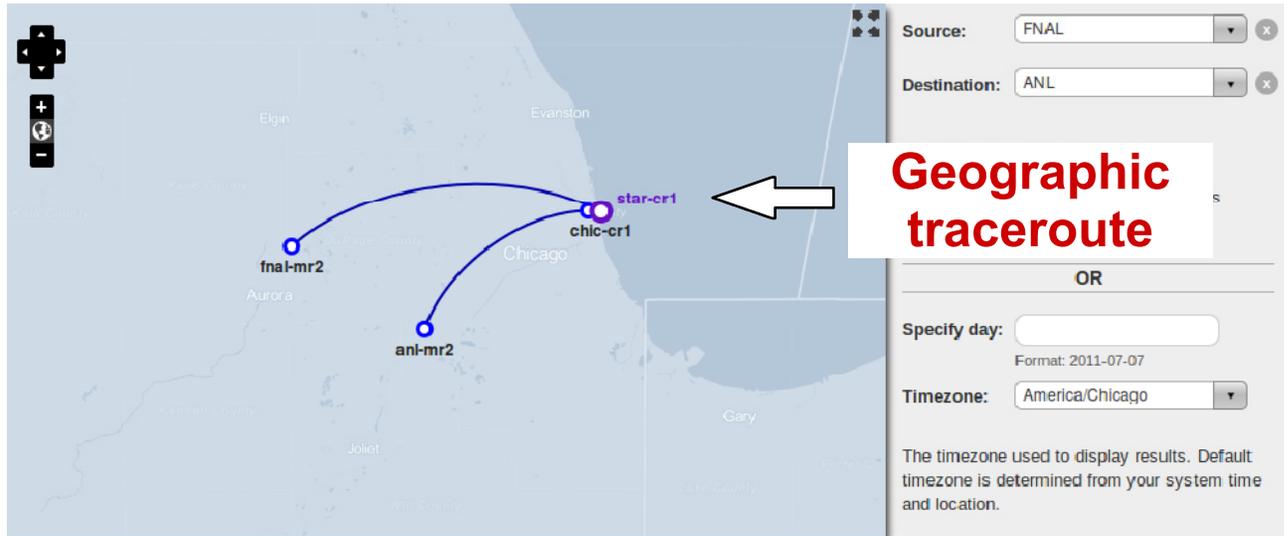
Last: 4 weeks 6 hours ago

### Site activity

Activity on the site and in your groups



# Front-end: Weathermap Results



**Path utilization chart**

**FNAL (fnal-mr2) to ANL (ani-mr2)**  
Jul 7, 2011 08:40 - Jul 7, 2011 11:40

- Create issue from these results
- Permanent link (bookmark and share results)
- <http://131.225.82.131/center/network?src=hub%3AFNAL&dst=>

**Create issue from query**

**Permanent link to query**



# Front-end: Weathermap Results (II)

## End to end results

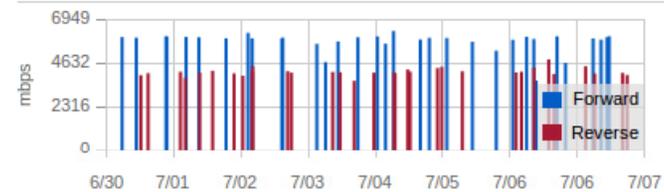
▸ Anomaly detection settings

Detect anomalies

Measurement	Average	Max	Min
Forward Throughput	5778mbps	6317mbps	3690mbps
Reverse Throughput	4157mbps	4810mbps	3678mbps
Forward Duplicates	0	0	0
Reverse Duplicates	0	0	0
Forward Loss	0.0	0.0	0.0
Reverse Loss	0.0	0.0	0.0
Forward Max delay	85ms	531ms	2ms
Reverse Max delay	88ms	476ms	2ms
Forward Min delay	1ms	2ms	1ms
Reverse Min delay	2ms	2ms	1ms

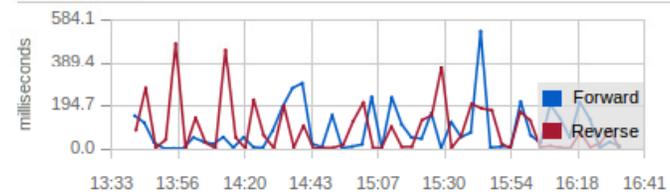
### Throughput

Jun 30, 2011 21:26 - Jul 7, 2011 13:24



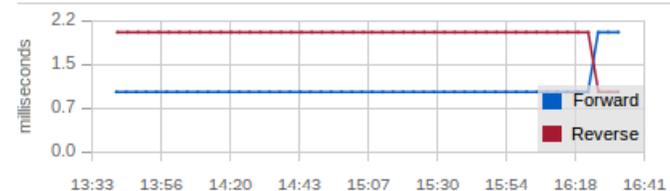
### Max delay

Jul 7, 2011 13:42 - Jul 7, 2011 16:32



### Min delay

Jul 7, 2011 13:42 - Jul 7, 2011 16:32



# Front-end: Weathermap Results

▼ Anomaly detection settings

Algorithm:

Sensitivity:

Elevation 1:

Elevation 2:

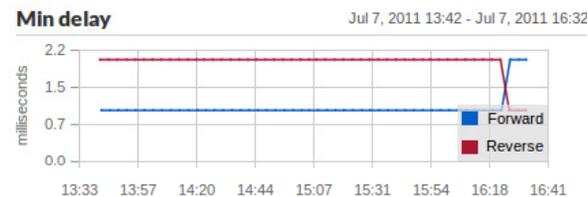
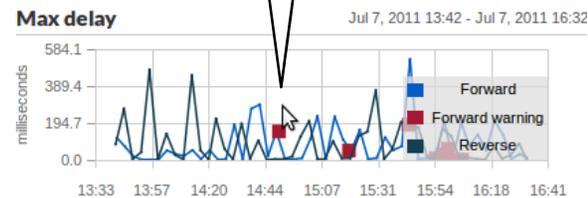
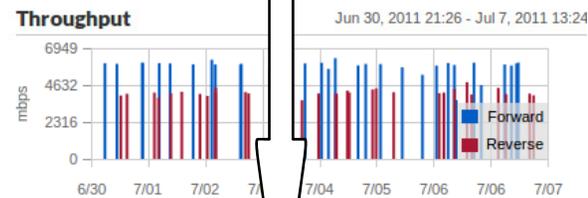
Window count:

Detect anomalies

← Experimental anomaly detection

## Anomalies shown on charts

Measurement	Average	Max	Min
Forward Throughput	5778mbps	6317mbps	3690mbps
Reverse Throughput	4157mbps	4810mbps	3678mbps
Forward Duplicates	0	0	0
Reverse Duplicates	0	0	0
Forward Loss	0.0	0.0	0.0
Reverse Loss	0.0	0.0	0.0
Forward Max delay	82ms	531ms	2ms
Reverse Max delay	88ms	476ms	2ms
Forward Min delay	1ms	2ms	1ms
Reverse Min delay	2ms	2ms	1ms



---

# Front-end: Easy Content Creation

- Rich text editor
- WYSIWYG and Markdown modes
- Embedded media support (YouTube, Vimeo)
- Code/syntax highlighting
- Powerful, secure, nice to use
- Email integration: custom notifications, replied by-email functionality



# Front-end: Measurement mesh health

- Shows which perfSONAR data is being collected by DRS
- Helps identify/troubleshoot problems in measurement infrastructure

## Hub status

Hub	snmp	bwctl	traceroute	owamp	pinger
SLAC	■	■	■	■	■
FNAL	■	■	■	■	■
JGI	■	■	■	■	■
BNL	■	■	■	■	■
NERSC	■	■	■	■	■
SNLL	■	■	■	■	■
LBL	■	■	■	■	■
LLNL	■	■	■	■	■
PNWG	■	■	■	■	■
ANL	■	■	■	■	■
PPPL	■	■	■	■	■
ORNL	■	■	■	■	■

Metadata 87  
Cached data 9584



# Front-end: End-to-end UI & data visualization

- **Streamlined UI:**

Reduce clicking and "knob turning" for common cases (inspired by Google instant).

- **Provide many views of data:**

Subway map, geographical map, and end-to-end tables provide several perspectives on path data. Site-centric and measurement health status provide logical views of site status. Chart data magnifying glass feature.

- **Self-revealing:**

Tool-tips integrated throughout UI provide users with context for interpreting results.

- **Highlight critical information:**

Chart highlighting technology attempts to make finding and isolating abnormal results easy.



---

# Dynamic Circuits Monitoring

- **Real-Time ESnet OSCAR circuits data:**
  - .Circuits Utilization is gathered in real-time, without caching
- **10 min delay for circuits pulling:**
  - .Circuits metadata pulled from ESnet topology service every 10 minutes to provide timely updates on active circuits

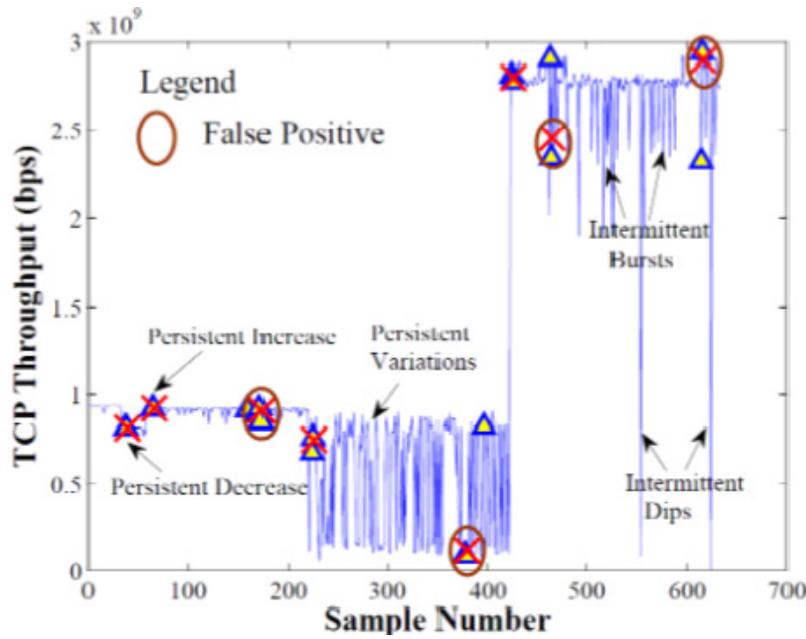


# ADS: Anomalies Detection Service

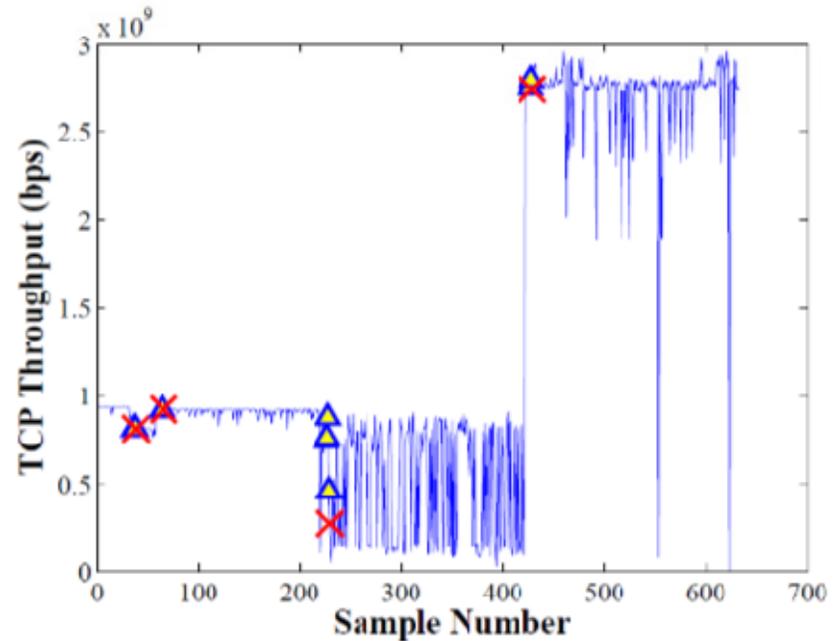
- Developed by Prasad Calyam (OSC)
- **REST** based web service, can be located anywhere
- Designed to work with any E-Center supported data sample or may send request to the DRS for the time series dataset
- Implements Static (**SPD**) or Adaptive Plateau Detection Algorithms (**APD**):
- Protocol is documented at:  
[https://cdcvs.fnal.gov/redmine/projects/ecenter/wiki/Anomalies\\_Detection\\_Service\(ADS\)](https://cdcvs.fnal.gov/redmine/projects/ecenter/wiki/Anomalies_Detection_Service(ADS))



# ADS: SPD and APD samples



**Static Sensitivity in SPD Scheme**

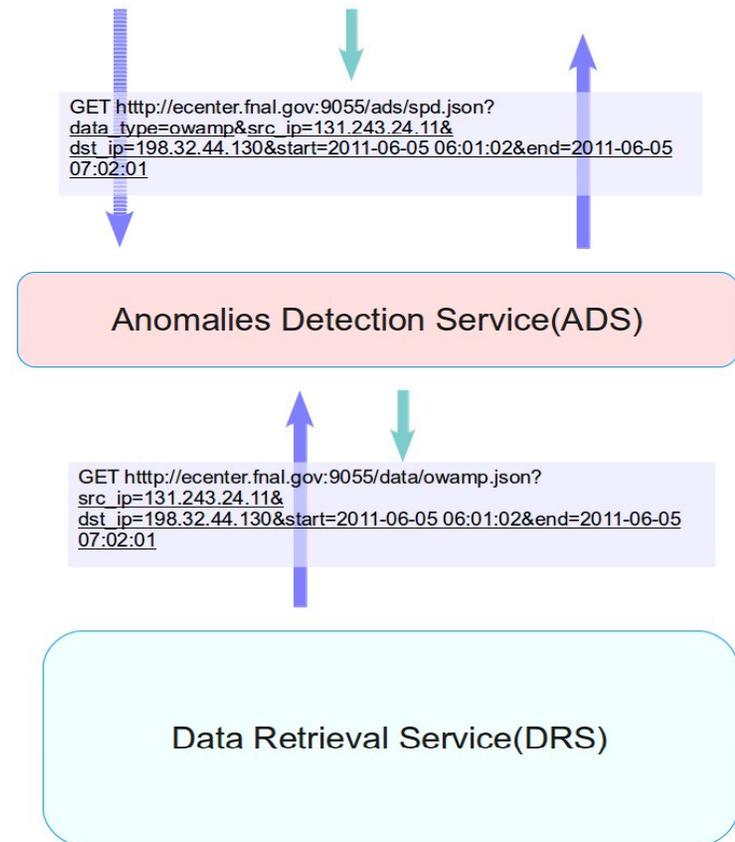


**Dynamic Sensitivity in APD Scheme**



# ADS: Integration with DRS

- ADS is standalone web service
- Can return results for supplied data sample
- OR, can send request to DRS for specific data type for time period
- Returns all warnings & critical anomalies found according to plateau detection parameters



---

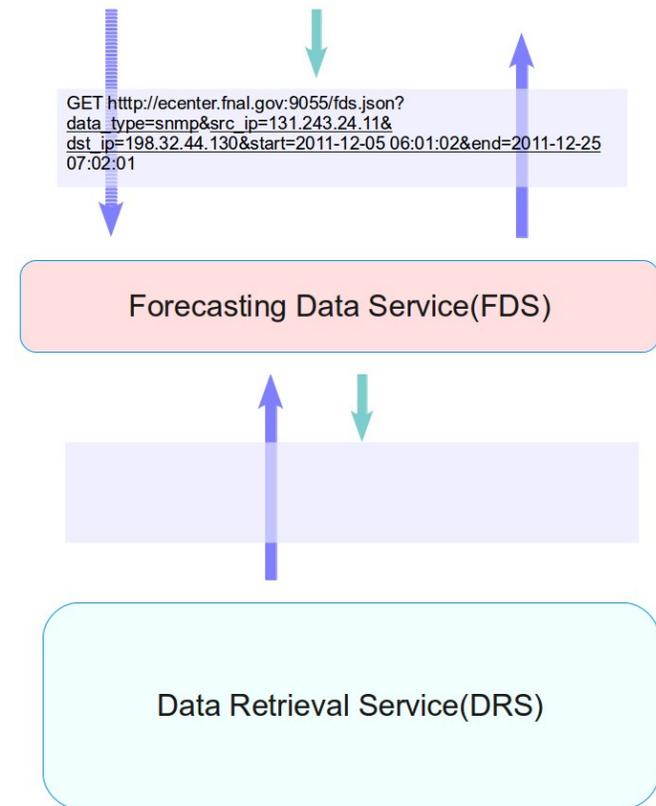
# FDS: Forecasting Data Service

- Based on Forecasting algorithm developed by Michael Frey at Bucknell University
- **REST** based web service, can be located anywhere
- Designed to work with any E-Center supported data sample or may send request to the DRS for the time series dataset
- Protocol is documented at:  
[https://cdcv.s.fnal.gov/redmine/projects/ecenter/wiki/Forecasting\\_Data\\_Service\(FDS\)](https://cdcv.s.fnal.gov/redmine/projects/ecenter/wiki/Forecasting_Data_Service(FDS))

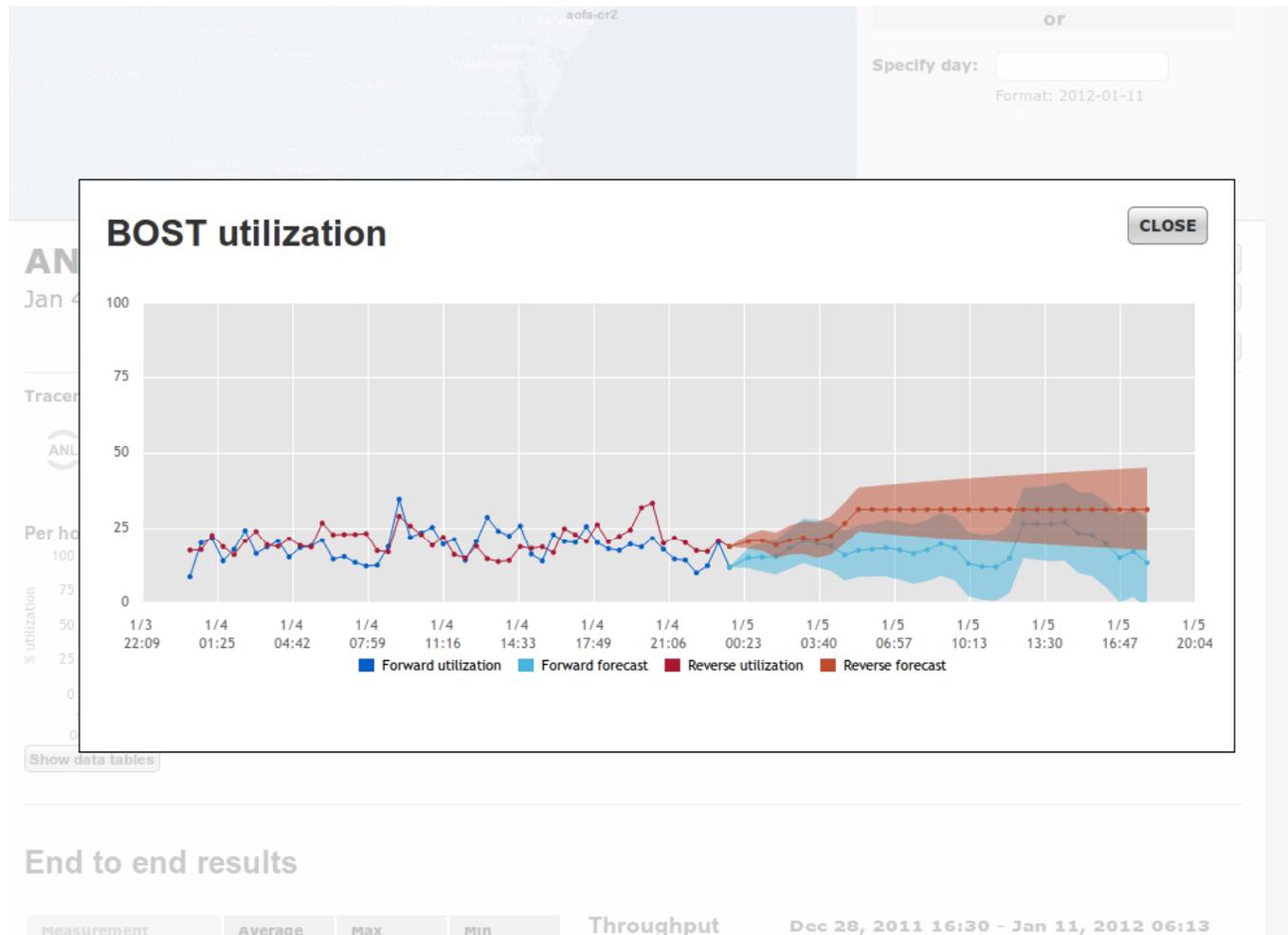


# FDS: Integration with DRS

- FDS is standalone web service
- It sends request to DRS for specific data type for some time period but for not less 200 data points – analysis limit
- Returns forecasted values with error bands



# FDS: In Action



# Forecasting UI



or

Specify day:   
Format: 2012-01-11

## ANL to BNL

Jan 4, 2012 00:00 to Jan 5, 2012 00:00

<http://localhost/center/ixTa> [Link](#)

Timezone: America/Chicago ▼

Type: SNMP ▼

[Get forecast](#)

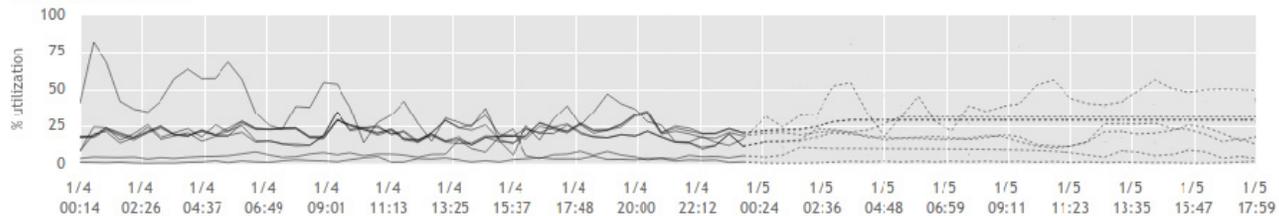
▶ [Anomaly settings](#)

[Detect anomalies](#)

Traceroute (logical) ■ Forward ■ Reverse [Click nodes to see detail](#)



### Per hop utilization



[Show data tables](#)

### End to end results

Measurement	Average	Max	Min
Forward Throughput	4080mbps	4606mbps	2209mbps

Throughput

Dec 28, 2011 16:30 - Jan 11, 2012 06:13



---

# ESCPS Extra Slides



---

# Basic ESCPS Functions

- High level: provide a network service to facilitate site use of circuit services
- Accept and process user/app requests for circuit services
- Provide local interface to & coordination of WAN circuit services
- **Configure local network infrastructure for use of circuits**
  - FNAL's contribution to the project...
- Monitor local network segments of end-to-end path



---

# Elements of ESCPS Model

- Aggregated Flow Endpoints (AFEs)
  - Sinks/sources for data flows; often clusters of systems
- Circuits
  - OSCARS constructs with L2/L3 terminations
- Virtual Paths
  - Complete end-to-end path between AFEs
- Rules:
  - configuration units that need to be deployed to extend a circuit to become a desired virtual path



---

# Basic LDC Rules

- L3-interface
- ACL
- Vlan
- QoS Class, QoS policer
- Route-map (Policy based mapping)

Vendor-specific syntax of these rules hidden in templates that can be selected/modified when creating a site's specific description of the rules. Easily modifiable for site's specifics.



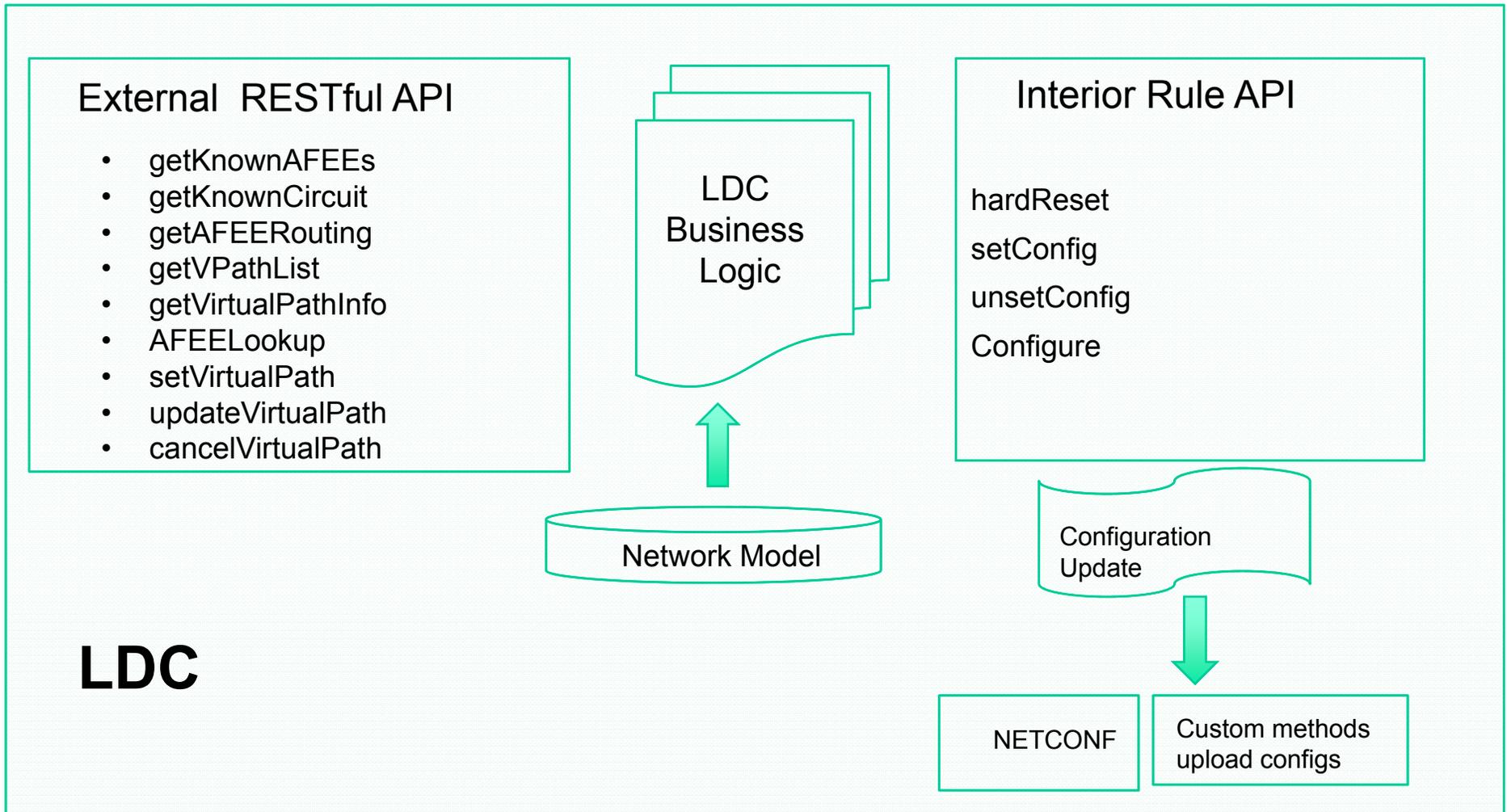
---

# LDC Rules Operations

- Unified operations to assemble rules in a virtual path across LAN:
  - ❑ hardReset – re-create rule from scratch
  - ❑ softReset – bring a rule into initial state, similar to above but does it less disruptive (depend on a rule)
  - ❑ setConfig – create a configuration fragment for rule ready to be uploaded in device(device(s))
  - ❑ unsetConfig – create a configuration fragment to undo setConfig
  - ❑ Configure – configure selected properties of rule (manipulate by rule's representation in server's memory and rule's operation data)

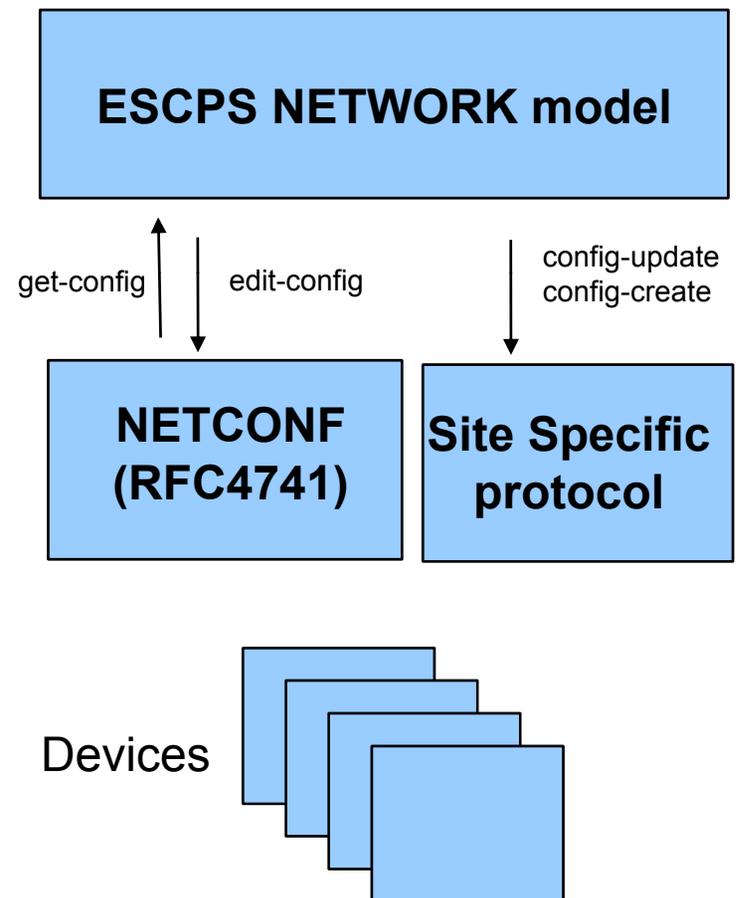


# Glue all together



# LDC Evaluations with Device Configuration Standards

- NetConf:
  - ❑ IETF Network Management Protocol (RFC4741)
  - ❑ RPC-based device configuration
  - ❑ Uses XML
- Limited vendor support
- Completed evaluation(s) on Cisco Nexus:
  - ❑ Complicated, complex, & somewhat confusing
  - ❑ Proprietary configuration methods still much easier



---

# ESCPS Project Short Term Objectives

- Complete evaluation of synergy w/ OSCARS platform:
  - Make use of OSCARS code where easy & beneficial
  - Determine longer term convergence path
- Integrate component work into functional prototype
  - Looking at ~6 month development effort
  - Means most ESCPS functions will remain outside of OSCARS
  - Seek out opportunities for collaboration (DYNES?)
- Get test-bed fully deployed



---

# Long Term Objectives (Time Permitting)

- Incremental Steps toward the goal of Integration of OSCARS capabilities into ESCPS.
- Items categorized as “Research and Development” along with OSCARS.
  - Topology System Management: Use the PerfSONAR to establish local topology, i.e. Topology Schema and Annotation.
  - Add AFE into Topology Schema and Store
    - It has the commonality with OSCARS’s new paradigm of Multi-Point to Multi-Point
  - LDC Functionalities added into OSCARS framework
  - Authentication in SVCMGR



---

# Long Term Vision beyond Current Project

- Integrate OSCARS capabilities into ESCPS
  - OSCARS is designed for WANs with well defined set of network technologies
  - Significant modifications are necessary to handle heterogeneous, hybrid end-site technologies, including L2, L3, DiffServ, IntServ
  - Extend the OSCARS code base by modifying and/or replacing components to add end-site functionality
- ESCPS becomes the end-site version of OSCARS
- Utilization of OpenFlow framework for device configuration
- Large scale deployment into DOE sites
- Network Resource Management & Co-scheduling along with CPU and Data Management



---

# OpenFlow

- OpenFlow support not part of ESCPS project plan
  - However, emergence of OpenFlow can't be ignored
- Current thinking = LDC OpenFlow support would entail just another rules set
  - May entail some modifications to API calls/handling as well
- Will try to evaluate OpenFlow over next several months
  - Independent of ESCPS project (and effort...)
  - But will be keeping OpenFlow capabilities in mind for ESCPS



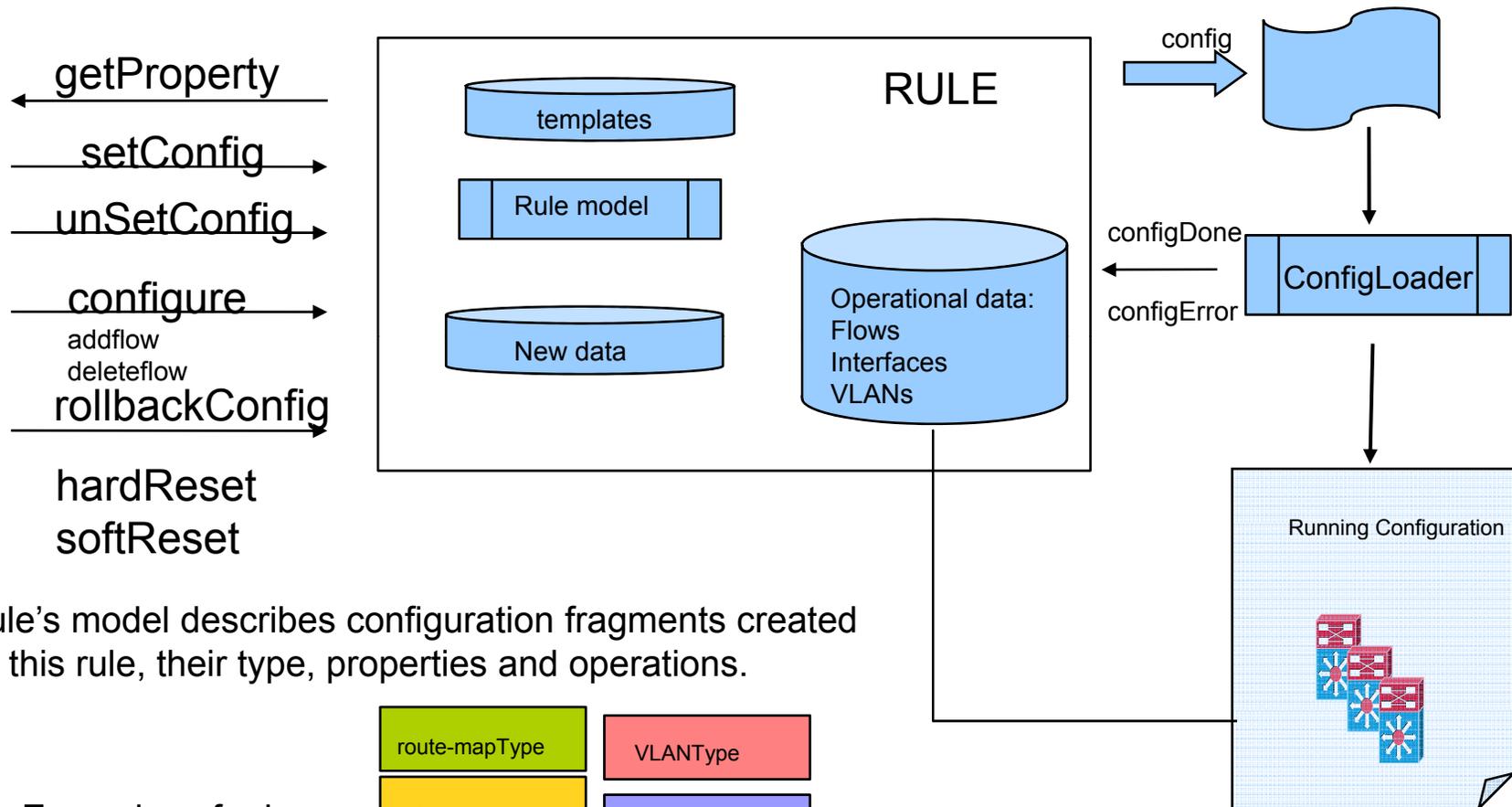
---

# RESTFuI LDC API

- getKnownAFEEs
- getKnownCircuit
- getAFEERouting
- getVPathList
- getVirtualPathInfo
- AFEELookup
- setVirtualPath
- updateVirtualPath
- cancelVirtualPath

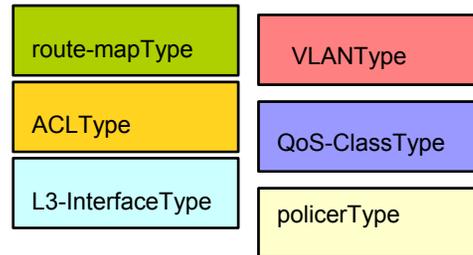


# Architecture of a Simple Type Rule



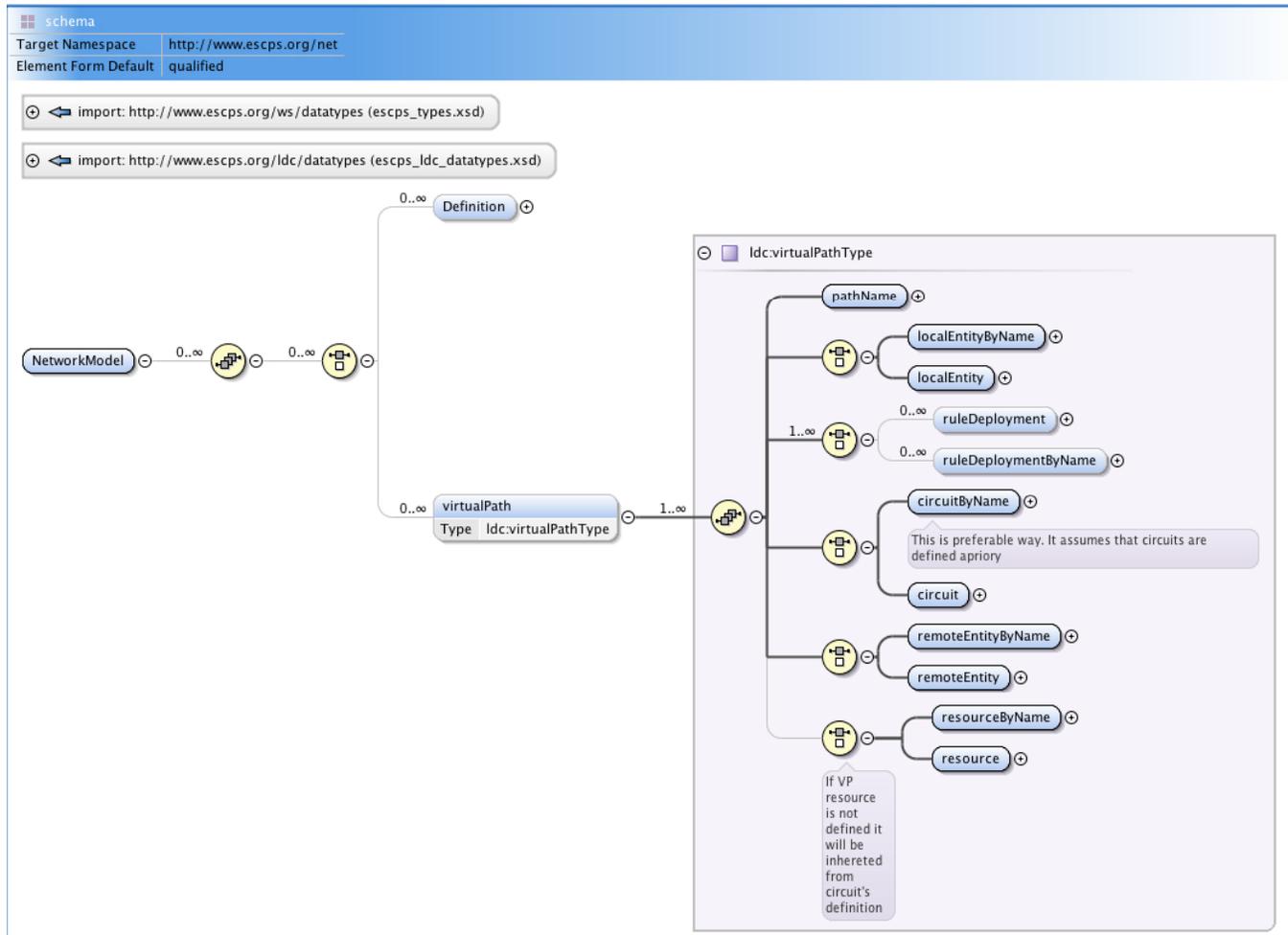
Rule's model describes configuration fragments created by this rule, their type, properties and operations.

Examples of rules



# Network Model Schema

Main abstract elements: Virtual Path, Flow Entity (AFEE), circuit, Rule, Resource)



## Network Model (Cont)

Defines a virtual path (paths) across LAN in terms of flow entities (AFEE), local and remote, and configuration rules:

..... A fragment of network model:

```
<virtualPath>
  <pathName xmlns="">UNL-DYNAMIC-VP-2</pathName>
  <localEntity xmlns="">
    <name>USCMS-T1-SRM</name>
    <address>131.225.204.1/32</address>
  </localEntity>
  <ruleDeploymentByName xmlns="">cms-wg-pbr-out</ruleDeploymentByName>
  <ruleDeploymentByName xmlns="">unl-circuit-dynamic</ruleDeploymentByName>
  <ruleDeploymentByName xmlns="">e2e-pbr-in</ruleDeploymentByName>
  <circuitByName xmlns="">unl-fnal-circuit-dynamic</circuitByName>
  <remoteEntity xmlns="">
    <name>UNL</name>
    <address>129.128.1.1/32</address>
  </remoteEntity>
</virtualPath>
</NetworkModel>
```



# Monitoring status an LDC service ticket...

**Display Resource Details**

Id: 675 Remote Ticket Id: 315  
 Status: Started Remote Status:  
 Application Class: Transactional Requestor: test  
 Bandwith In: Bandwith Out:  
 Total Bandwith:  
 Start Time: 06/23/2011 04:42 PM End Time:  
 Duration: 8280000

Ticket Workflow					
Status	Time	Message	Event	Process	Event Result
Remote Booking	06/23/2011 04:45:12 PM	Request from service	Remote Ticket Booking		Recieved
Local Booking	06/23/2011 04:45:41 PM	Process Flow Started	Resource Bandwidth Reservation	resourceManager	Recieved
			Resource Bandwidth Reservation	secondProcess	-
Booked	06/23/2011 04:46:03 PM	Resource Manager booked Ticket			Recieved
Reserved	06/23/2011 04:46:10 PM	Ticket is consolidated			Recieved
Started	06/23/2011 04:46:10 PM	Ticket Start Time reached			Recieved

Local Flow Entity

Name:	USCMS-T1
Addresses	
Applications	

Remote Flow Entity

Name:	kazim
Addresses	
Applications	

Close



# Monitoring A Reservation...

**Display Resource Details**

Id: 692      Remote Ticket Id:  
 Status: Active      Remote Status:  
 Application Class: Default      Requestor:  
 Bandwith In:      Bandwith Out:  
 Total Bandwith:  
 Start Time: 06/23/2011 04:42 PM      End Time: 10/14/2014 12:42 AM  
 Duration:

Reservation Workflow					
Status	Time	Message	Event	Process	Event Result
Accepted	06/23/2011 04:45:40 PM	New reservation created			Recieved
Activating	06/23/2011 04:45:41 PM	Process Flow Started	Network Config Setup:activating	Idc	Recieved
			Configuring DCN:DCNConfigs	Idc	Received
Active	06/23/2011 04:46:21 PM	LDC activated reservation			Recieved

Local Flow Entity	
Name:	USCMS-T1
Addresses	
Applications	

Remote Flow Entity	
Name:	kazim
Addresses	
Applications	

Close



# Monitoring ticket's status ....

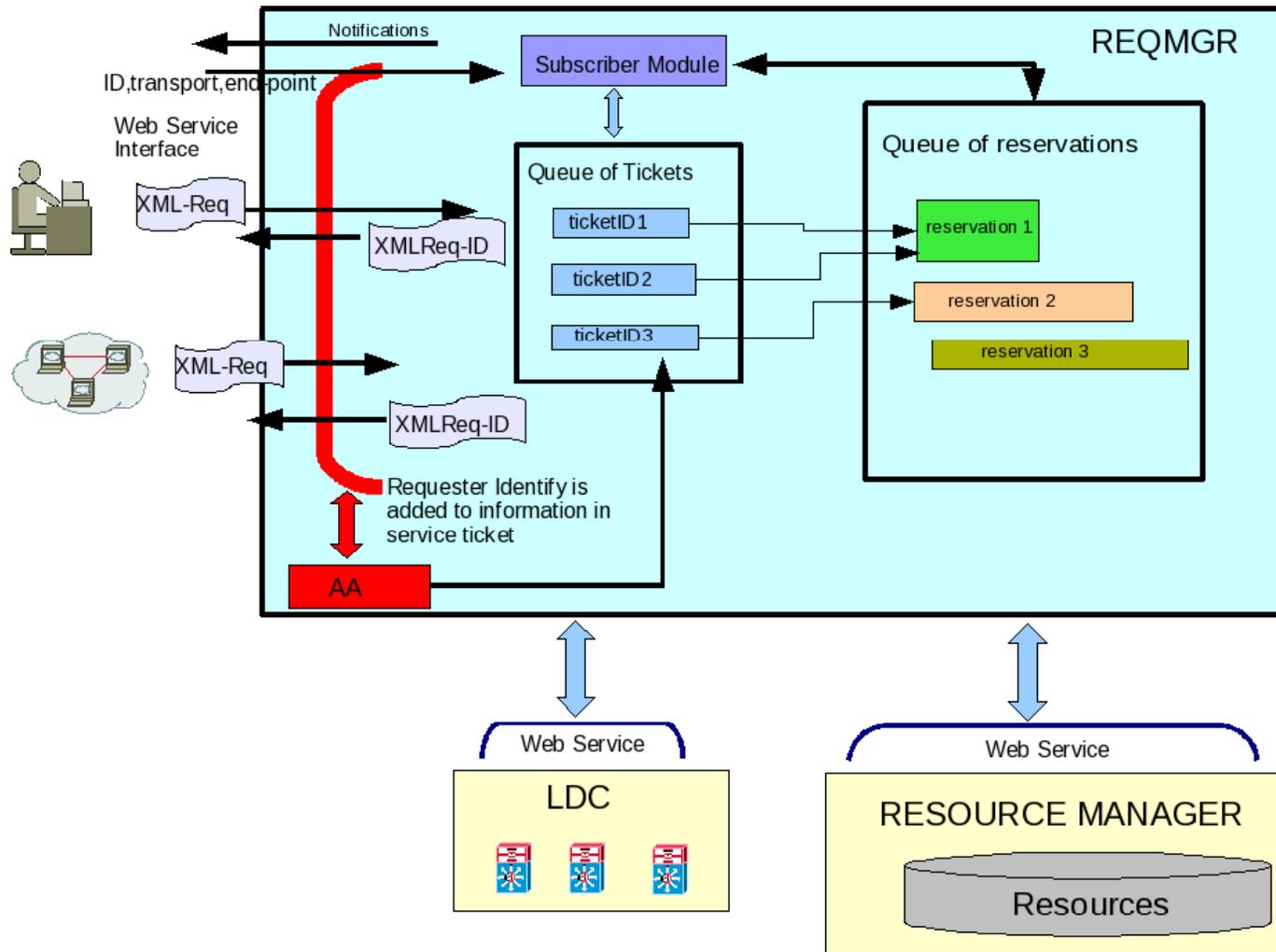
Active Tickets													
Ticket Id	Reservation Id	Ticket Status	Reservation Status	Remote Ticket Id	Remote Ticket Status	Local Entity	Remote Entity	Application Class	Start Time	End Time	Requestor	Chart	Cancel
667	688	Started	Active	16		USCMS-T1	kazim	Streaming	06/23/11 16:41:41	11/09/13 02:21:41	test	Display	Cancel
668	689	Started	Active	67		USCMS-T1	kazim	Real-Time	06/23/11 16:42:00	12/12/13 10:22:00	test	Display	Cancel
669	690	Started	Active	79		USCMS-T1	kazim	Bulk Scavenger	06/23/11 16:42:05	04/27/12 17:22:05	test	Display	Cancel
670	690	Started	Active	23		USCMS-T1	kazim	Bulk Scavenger	06/23/11 16:42:13	03/04/13 02:22:13	test	Display	Cancel
671	691	Started	Active	98		USCMS-T1	kazim	Data Movement	06/23/11 16:42:13	02/02/13 22:22:13	test	Display	Cancel
672	692	Started	Active	14		USCMS-T1	kazim	Default	06/23/11 16:42:16	10/14/14 00:42:16	test	Display	Cancel
673	692	Started	Active	163		USCMS-T1	kazim	Default	06/23/11 16:42:23	04/24/12 06:02:23	test	Display	Cancel
674	692	Started	Active	42		USCMS-T1	kazim	Default	06/23/11 16:42:26	04/03/14 23:22:26	test	Display	Cancel
675	692	Started	Active	315		USCMS-T1	kazim	Transactional	06/23/11 16:42:51	09/27/11 12:42:51	test	Display	Cancel
676	689	Started	Active	223		USCMS-T1	kazim	Real-Time	06/23/11 16:43:15	09/23/11 08:43:15	test	Display	Cancel

[<](#)
[<<](#)
[<](#)
[1](#)
[2](#)
[3](#)
[4](#)
[5](#)
[6](#)
[7](#)
[8](#)
[9](#)
[>](#)
[>>](#)
[>](#)

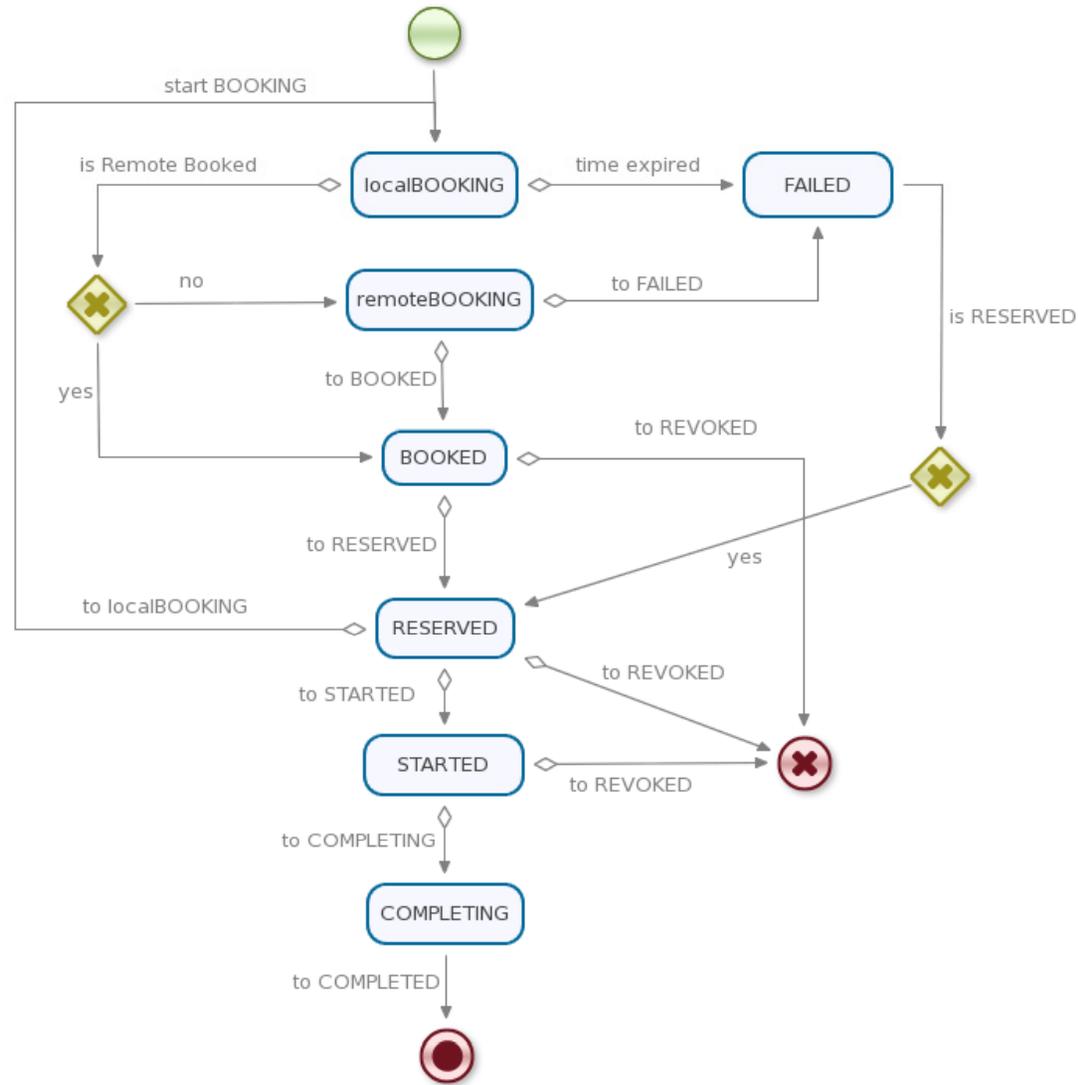
118 tickets found, displaying 10 tickets, from 1 to 10. Page 1 / 12



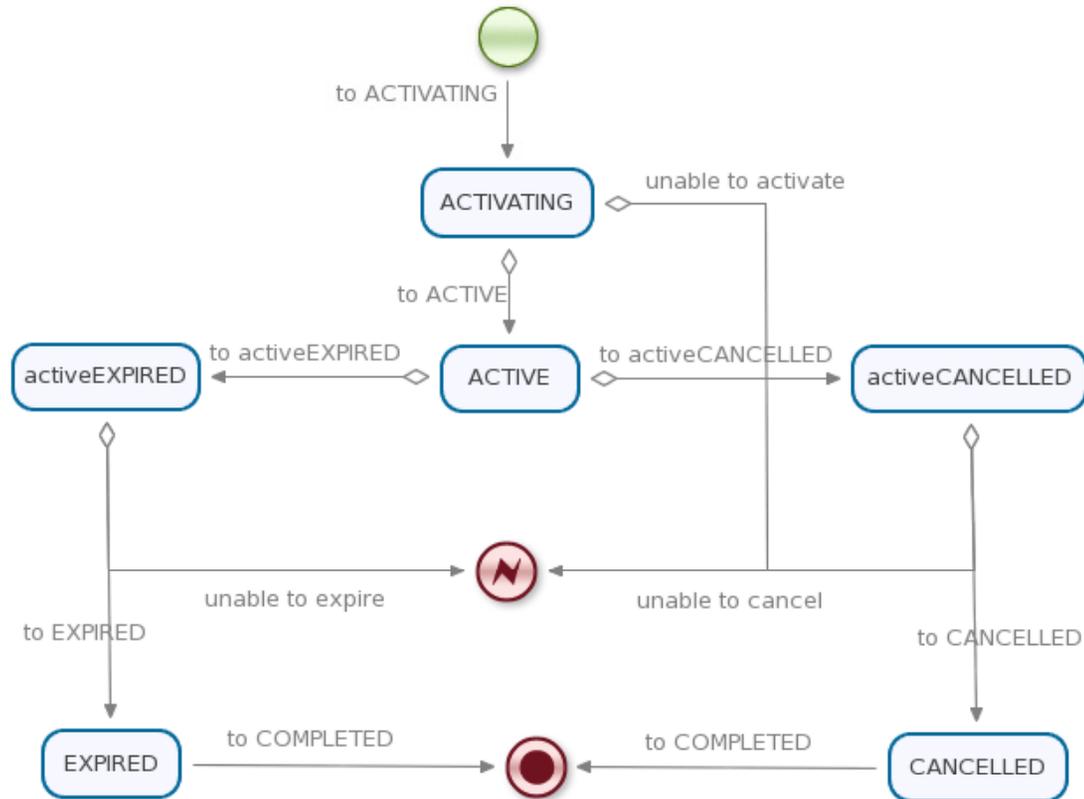
# LDC Service Manager Architecture



# Simplified LDC Ticket's workflow



# Simplified LDC Reservation's workflow



---

# LDC Service Manager RESTFul API

getServiceTicket

searchServiceTicket

addServiceTicket

updateServiceTicket

getTicketStatus

updateTicketStatus

cancelTicket

listReservation

getReservationInfo

- getCircuits
- getCircuitInfo
- getListAFEEs
- afeeLookup



# Signaling from additional external processes can be to Ticket's workflow via XML configuration file (needs server's restart)

```
<events xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="file:events.xsd">
  <global>
    <processId name="resourceManager">
      <userid>resID</userid>
      <password>asasa</password>
    </processId>
    <processId name="secondProcess">
      <userid>secondProcessID</userid>
      <password>asasa</password>
    </processId>
    <processId name="ldc">
      <userid>ldcUser</userid>
      <password>asasa</password>
    </processId>
    <event>
      <timeout>300000</timeout>
    </event>
  </global>
```

```
<state name="ACTIVATING" check="ALL">
  <event id="activating">
    <name>Network Config
Setup:activating</name>
    <timeout>250000</timeout>
    <process id="ldc">
      <defaultMessage>LDC
activated reservation</defaultMessage>
    </process>
    <defaultErrorMessage>Resource Manager unable
to book Ticket</defaultErrorMessage>
  </event>
  <event id="DCNConfig">
    <name>Configuring
DCN:DCNConfigs</name>
    <process id="ldc">
      <defaultMessage>LDC
activated reservation</defaultMessage>
    </process>
    <defaultErrorMessage>Resource Manager unable
to book Ticket</defaultErrorMessage>
  </event>
</state>
```



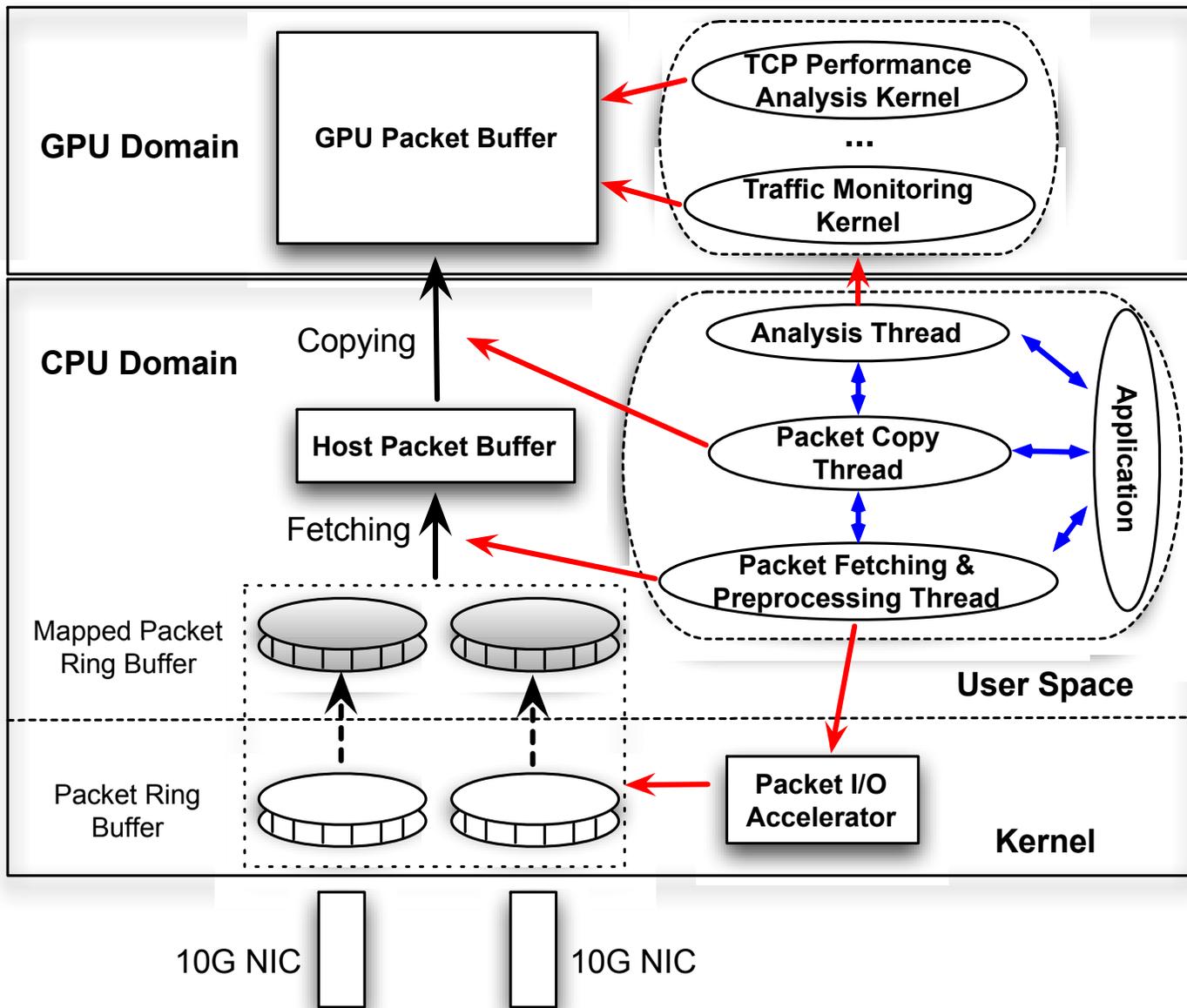
</global>



---

## Other Network R&D Efforts

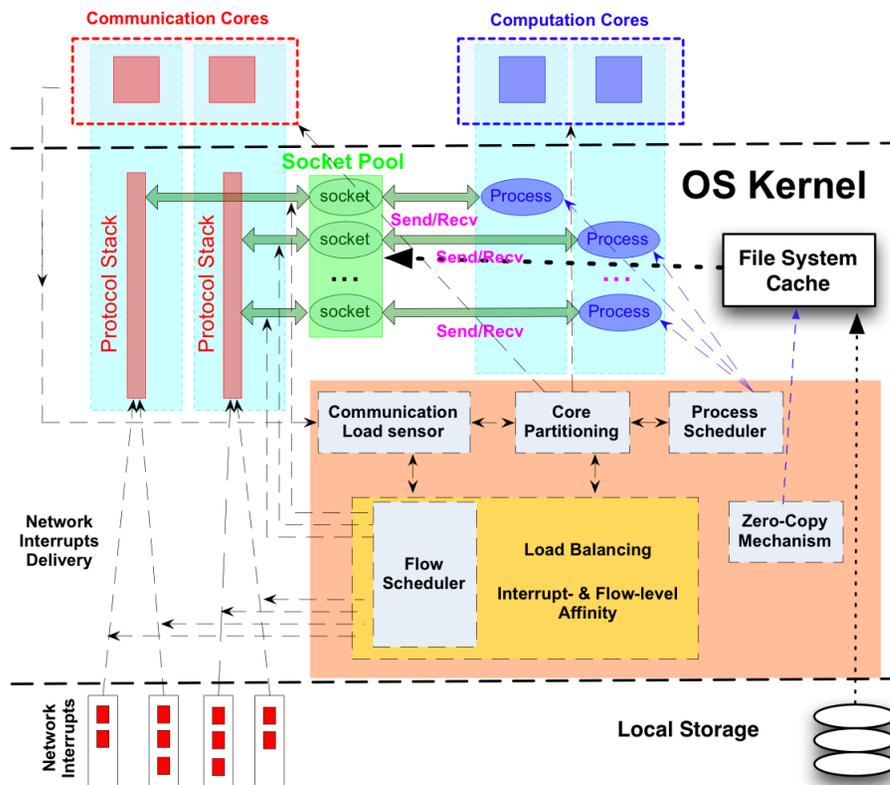




G-Debugger



# Multicore Communication Architecture (MCA) design

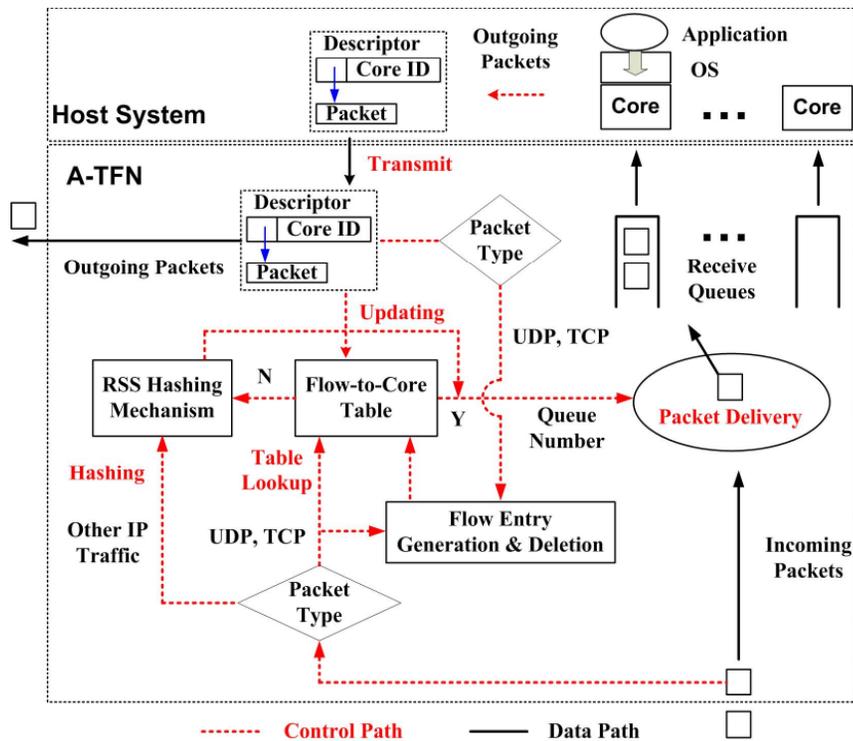


From the functionality point of view, MCA has three components: the host protocol stack, network application interfaces, operating system supports.



# A-TFN Mechanism

- A-TFN steers incoming network data to the same core on which its application thread resides and ensures in-order packet delivery.



- For transport layer traffic, A-TFN supports a Flow-to-Core table to maintain the relationship “Traffic Flows → Cores (Applications)” in the NIC, with one entry per flow. Each entry tracks which core a flow should be assigned to.
  - A-TFN applies a very simple yet effective mechanism to update the Flow-to-Core table, which identifies the core on which the application is running faster and more accurately than Flow Director.
  - A-TFN deals with process migration and has a mechanism to ensure in-order packet delivery.

