ANSE: Advanced Network Services for [LHC] Experiments

Artur Barczyk, California Institute of Technology
for the ANSE team

LHCONE Point-to-Point Service Workshop

Geneva, December 13th, 2012

Introduction to ANSE

- A project funded by NSF CC-NIE program
  - Campus Cyberinfrastructure
  - ➡ Network Infrastructure and Engineering

    Advanced network methods aimed at serving major science projects

    LHC (ATLAS and CMS) clearly qualify;

    NSF: Emphasis on universities hence on Tier2 and Tier3s Operations
- Two years of funding, official starting date Jan 2013, ~3 FTEs
- PIs: Harvey Newman, PI, Caltech

  Shawn McKee, co-PI, University of Michigan

  Paul Sheldon, co-PI, Vanderbilt University

  Kaushik De, co-PI, University of Texas at Arlington

  Artur Barczyk, co-PI, Caltech

Objectives and Approach

- **Deterministic, optimized workflow is the goal**
  - **Use network resource allocation along with storage and CPU resource allocation in planning data and job placement**
  - **Improve overall throughput and task times to completion**
- **Integrate advanced network-aware tools in the mainstream production workflows of ATLAS and CMS**
  - **use tools and deployed installations where they exist**
    - i.e. build on previous manpower investment in R&E networks
  - **extend functionality of the tools to match experiments' needs**
  - **identify and develop tools and interfaces where they are missing**
- **Green-Field, but not Terraforming**
  - **Introduce new/recent concepts**
  - **Build on several years of invested manpower, tools and ideas (some since the MONARC era)**

Methodology

- **Use agile, managed bandwidth for tasks with levels of priority along with CPU and disk storage allocation.**
  - Allows one to define goals for time-to-completion, with reasonable chance of success
  - Allows one to define metrics of success, such as the rate of work completion with reasonable resource use
  - Allows one to define and achieve “consistent” workflow
- **Dynamic circuits a natural match**
  (as in DYNES for Tier2s and 3s)
- **Process-Oriented Approach**
  - Measure resource usage and job/task progress in real-time
  - If resource use or rate of progress is not as requested/planned, diagnose, analyze and decide if and when task replanning is needed
- **Classes of work: defined by resources required, estimated time to complete, priority, etc.**

Tool Categories

- **Monitoring**
  - **Allows reactive use – react to events or situations in the network**
    - throughput measurements; possible actions:
      - raise alarm and continue
      - abort/restart transfers
      - choose different source
    - topology monitoring; possible actions:
      - influence source selection
      - raise alarm (e.g. extreme cases like site isolation)
- **Network Control**
  - **Allows pro-active use**
    - reserve Bandwidth -> prioritize transfers, remote access flows, etc.
    - Co-scheduling of CPU, storage and network resources
    - create custom topologies -> optimize infrastructure to operational conditions
      - e.g. during LHC running period vs reconstruction/re-distribution

The Network API

- **The network APIs have been developed by "network folks"**
  - **not a critique, we needed a starting point!**
  - **Does it match what users need?**
- **Some ideas collected at PhEDEx wiki (thanks to T. Wildish)**
  - https://twiki.cern.ch/twiki/bin/view/CMS/PHEDEXSupportForDynamicCircuits
- **Q: Is the API provided (e.g. NSI-CS) adequate?**
  - **do we need to develop "bandwidth budget" scheme?**
  - **what happens when reservation request is denied?**
    - what information does the requesting app provide
      - start/end times?
      - strict on capacity? or duration?
      - or data set size? (can it be verified by the service provider? reliably?)
    - what information is returned
      - YES, NO, alternatives, …?
- **ANSE product could be the 'glue'**
  - **e.g. a library using NSI API as primitives**

CMS Example: Data Source Selection

- Close and active collaboration with PhEDEx team
  - Direct participation of Tony Wildish in ANSE
- Support decision on source location for replication
  - Aka “router” in PhEDEx
- Today uses past statistics to select the best source site for data transfers
- Recently hooks have been implemented to use external input as “router hints”.
- ANSE could expand this, using
  - topology description/monitoring information
  - perfsonar measurement data
  - circuit setup confirmation
  - …

ATLAS example:…

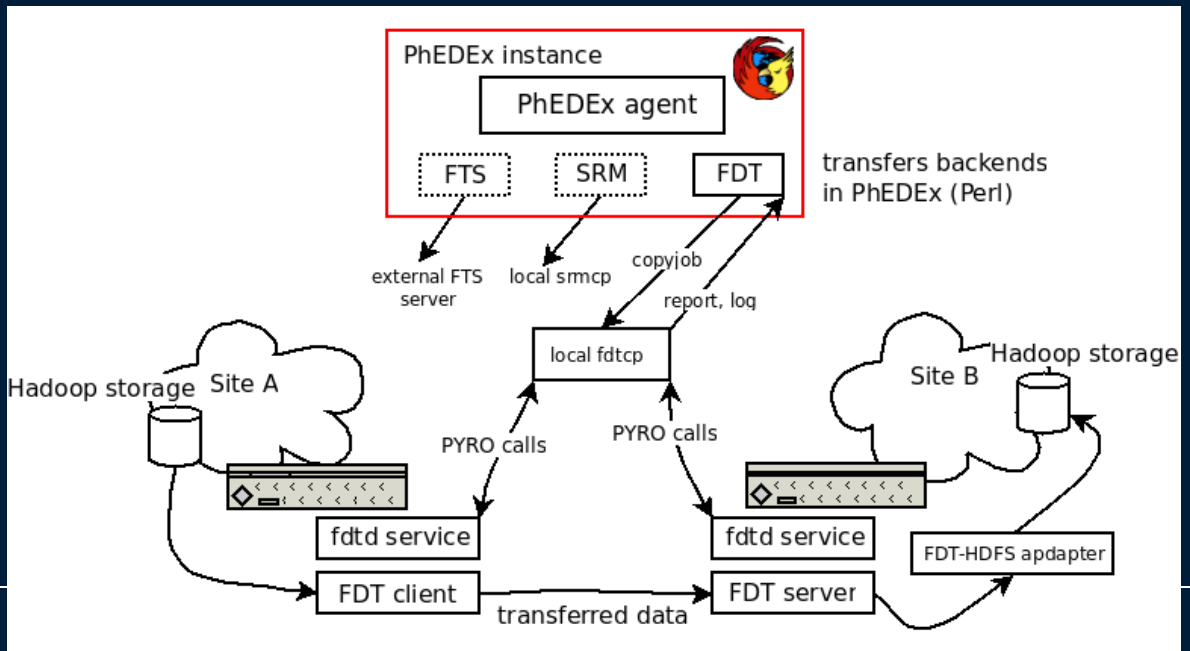
- …you've seen it all in Kaushik De's presentation earlier today

Relation to DYNES

- In brief, DYNES is an NSF funded project to deploy a 'cyberinstrument' linking ~40 US campuses through Internet2 dynamic circuit backbone
  - based on ION service, using OSCARS technology, see E. Boyd's slides
- DYNES instrument is intended as a production-grade 'starter-kit'
  - comes with a disk server, inter-domain controller (server) and FDT installation
  - FDT code includes OSCARS IDC API -> reserves bandwidth, and moves data through the created circuit
    - "Bandwidth on Demand", i.e. get it now or never
    - routed GPN as fallback
- The DYNES system is naturally capable of advance reservation
- All we need is the right agent code inside CMS/ATLAS to call the API whenever transfers involve two DYNES sites

DYNES/FDT/PhEDEx

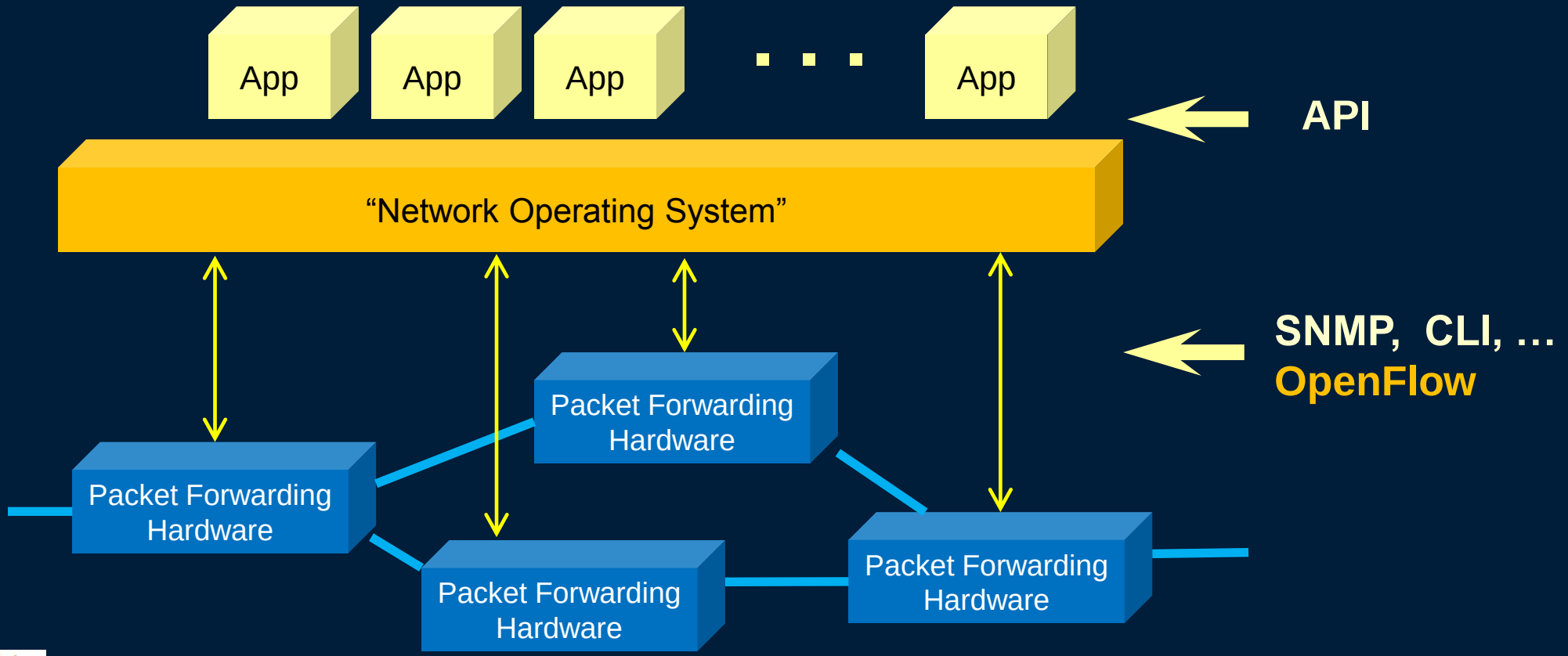
- FDT integrates OSCARS IDC API to reserve network capacity for data transfers
- FDT has been integrated with PhEDEx at the level of download agent
- Basic functionality OK
  - more work needed to understand performance issues with HDFS
- Interested sites are welcome to test
- With FDT deployed as part of DYNES, this makes one possible entry point for ANSE



- **Of course, the new kid on the block is…**
  (actually not even that new)

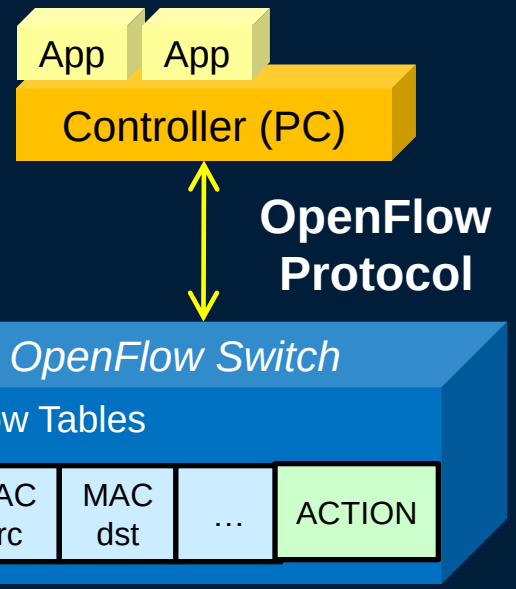
Software Defined Networking

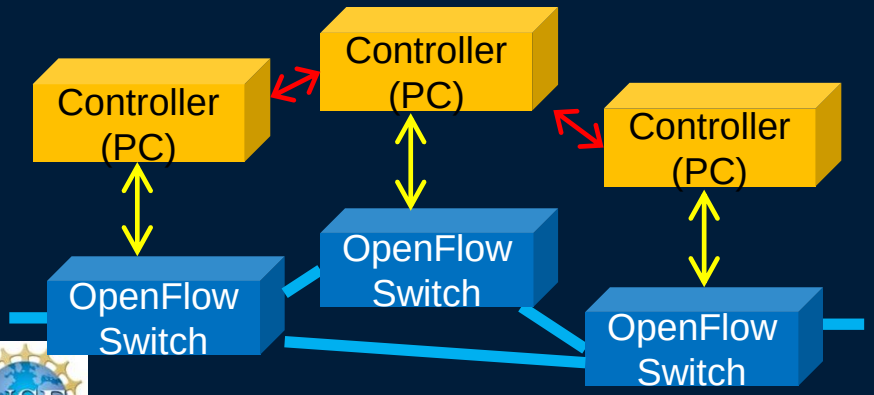
- **SDN Paradigm - Network control by applications; provide an API to externally define network functionality**

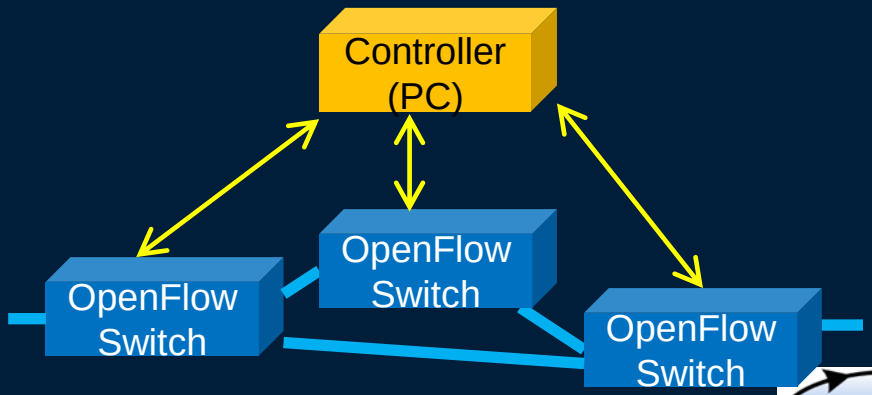


OpenFlow

- **Standardized SDN protocol**
  - **Open Networking Foundation (https://www.opennetworking.org/)**
- **Let external controller access/modify flow tables**
- **Allows separation of control plane and data forwarding**
- **Simple protocol, large application space**
  - **Forwarding, access control, filtering, topology segmentation, load balancing, …**
- **Distributed or centralized**
- **Reactive or pro-active**







so…

- **OpenFlow deployment is growing fast**
  - **in particular in virtualised data center environment**
  - **In the WAN - one example is Internet2's OS3E/NDDI network**

- **In LHCONE, we need to (continue to) investigate how OpenFlow is best used**
  - **this is done through one of the activities in the LHCONE Architecture WG**
    - One use case example is the WAN multipath fabric project, recently demonstrated by Caltech at SC'12

- **ANSE will follow the developments in LHCONE**

Summary

- ANSE project aims at integration of advanced network services in the LHC experiment's SW stacks
- Through interfaces to
  - Monitoring services (PerfSONAR-based, MonALISA)
  - Bandwidth reservation systems (through protocols like NSI and IDCP)

- Working with
  - PanDA system in ATLAS
  - PhEDEx in CMS

- The goal is to make deterministic workflows possible





QUESTIONS?

Artur.Barczyk@cern.ch