Design Patterns for Data-Intensive Science

Inder Monga
CTO, Energy Sciences Network
Lawrence Berkeley National Lab

INDIS workshop
SC15, Austin
Importance of Design Patterns

- Design pattern is a general reusable solution to a commonly occurring problem within a given context
  - Also defines interactions of objects at an abstract level
- The ‘Internet’ Design Pattern has served us well

- But has it solved all our problems?
When do you know that designs are rotting?

- Four characteristics*
  - **Rigidity**
    - Difficult to change
  - **Fragility**
    - Break at unexpected places, if change is made
  - **Immobility**
    - Inability to reuse, design anew
  - **Viscosity**
    - Hacks are easier to implement than proper changes

Experimental and observational science is at crossroads

- Data volumes are increasing faster than Moore’s Law
- New algorithms and methods for analyzing data
- Infeasible to put a supercomputing center at every experimental facility

What are the system design patterns for Data-Intensive Science?
Design Pattern #1: Science DMZ

Data Transfer Node
- High performance
- Configured for data transfer
- Proper tools

Engagement
- Partnerships
- Education & Consulting
- Resources & Knowledgebase

Dedicated Systems for Data Transfer

Performance Testing & Measurement

Network Architecture

Science DMZ
- Dedicated location for DTN
- Proper security
- Easy to deploy - no need to redesign the whole network

perfSONAR
- Enables fault isolation
- Verify correct operation
- Widely deployed in ESnet and other networks, as well as sites and facilities
#1: Science DMZ Design Pattern

![Diagram of Science DMZ Design Pattern with labels for Border Router, Enterprise Border Router/Firewall, Site/Campus LAN, Site/Campus Virtual Circuits, High performance Data Transfer Node, and perfSONAR.]
Design Pattern #2: End-to-End paths, across domains

Large Hadron Collider (LHC)
International Thermonuclear Experimental Reactor (ITER)
Genomics
ESnet
General Internet
#2: End-to-end paths

- Science Collaborations are global and NRENs local
- Instruments are distributed or located in single region, producing huge amounts of data
- ‘Elephant flows’ get their own traffic-engineered channel
There are some emerging design patterns...

- ...that need participation from the IT infrastructure and application community
Design Pattern #3: Superfacility
Vision: A network of connected facilities, software and expertise to enable new modes of discovery
#3: Superfacility Prototype and Use Case

Real-time analysis of ‘slot-die’ technique for printing organic photovoltaics, using ALS + NERSC (SPOT Suite for reduction, remeshing, analysis) + OLCF (HipGISAXS running on Titan w/ 8000 GPUs).


Results presented at March 2015 meeting of American Physical Society by Alex Hexemer. Additional DOE contributions: GLOBUS (ANL), CAMERA (Berkeley Lab)
There are some strong potential design patterns...

- ...that need research, development, brainstorming, and frankly, early adopters
Design Pattern #4: Network Operating Systems

• Networks have been managed as a set of discrete, autonomous entities sharing state with each other

• Pros
  – Resilience
  – Easy to grow by adding another autonomous entity

• Cons
  – Suboptimal resource allocation
  – Opaqueness

• How can we get the benefits of global knowledge while catering to multiple applications, and offering optimal resource allocation?
#4: Moving from Network MS to Network OS

Notional ESnet Network Operating System

- Federated applications
- ESnet Network Operating System (ENOS)
- SDN Controllers
- OSCARS
- Performance Agents
#4: ESnet Network Operating System (ENOS)

- Platform to expose network programmability to science applications
- Multi-domain VPN service over multi-continent testbed (DOE & Corsa Booth, SC15)
Design Pattern #5: Network Analytics

• Data being generated by the network but not analysed or available for real-time analysis
  – The ability to ask questions of historical network data, and get answers
  – The answers updated with new data in near real-time
  – SNMP data, Flow data, Topology data, etc..

• Architecture consists of three layers:
  – Batch processing: precomputing large amounts of data
  – Speed or real time: minimize latency by doing real time calculations
  – Layer to respond to queries: interfacing to provide the results
#5: Analytics Example: Lambda architecture
Leverage cloud computing tools to put together a pipeline

Generic design pattern picture – modified to work with specific cloud computing technologies
Summary

• Design patterns in network architecture and measurements is extremely important for scale

• Design patterns for networking should not stop at ‘connectivity and reachability’ – which is what the IPv4/v6 network provides us today

• Serving the Data Intensive Science needs has established new infrastructure design patterns, with a potential for lot more emerging

• What’s the role of SCinet and NRE?
Potential role of SCinet

• Help us **discover, implement, explore** and **operationalize** new infrastructure design patterns

• Create an environment where new ideas are overlaid over established best practices and tested in a safe manner

• Transition technologies quickly – feel free to engage vendors and open-source community

• Learn and re-invent!