The Elf On the Yellow Elephant

Internet2 Global Summit

Ze Yu@UF
The Yellow Elephant

- Big Data
- Hadoop – The Yellow Elephant
  - Standard tool to process Big Data
- Become the “OS” of the datacenter of Big Data
The First Principle of Hadoop – Locality

- Do not move data around
  - Since they are big…
- Instead, move computation to the data
Locality-Aware Scheduling

Scheduler

Rack 1
- Task 2
  - Slave
  - 2 3 6
- Task 1
  - Slave
  - 2 5 8
  - 1 2

Rack 2
- Task 3
  - Slave
  - 3 5 8
- Task 1
  - Slave
  - 1 4 9
  - 1 3

Rack 3
- Task 2
  - Slave
  - 1 7 9
- Task 3
  - Slave
  - 4 6 7
  - 2 3

Job1 ← File 1:
Job2 ← File 2:
Locality Is Hard To Get

- It is a network related property
- The datacenter network is a black box to the big data
  - Inter-datacenter locality-aware deployment is a mission impossible
- Currently, using manual configuration
Summary

• If we consider Hadoop as the “OS” of datacenter for Big Data, the network is the *missing piece* of this OS.
Elf

Using Software Defined Network to manage network for Big Data
Why SDN

• Programmable control
  – No more manual work

• Global view
  – Knows everything about the network

• Easy for integration
  – The rendezvous of compute & network

• Flexible
  – Handle the dynamics of the network
Elf - Example

Locality-Aware Computation

Where should I put task A?

Hadoop

Node M is the best place, N is #2

The Elf SDN Service
Elf - Example

Network-Driven Computation

Spare capacity between A → B, who needs it?

Job X could use it.

The Elf SDN Service
Elf - Example

Timely Data Prefetching

When should I copy data from a to b?

Hadoop

The Elf SDN Service

20 sec earlier
Takeaways

• Big data ecosystem relies more and more on being network-aware
• SDN can help Big Data as its network controlling knob
Thanks
My Wish List of SDN

- More performance-related control primitives added to SDN
  - Bandwidth control e.t.c
  - Both northbound and southbound API