T2_US_Nebraska Site Report

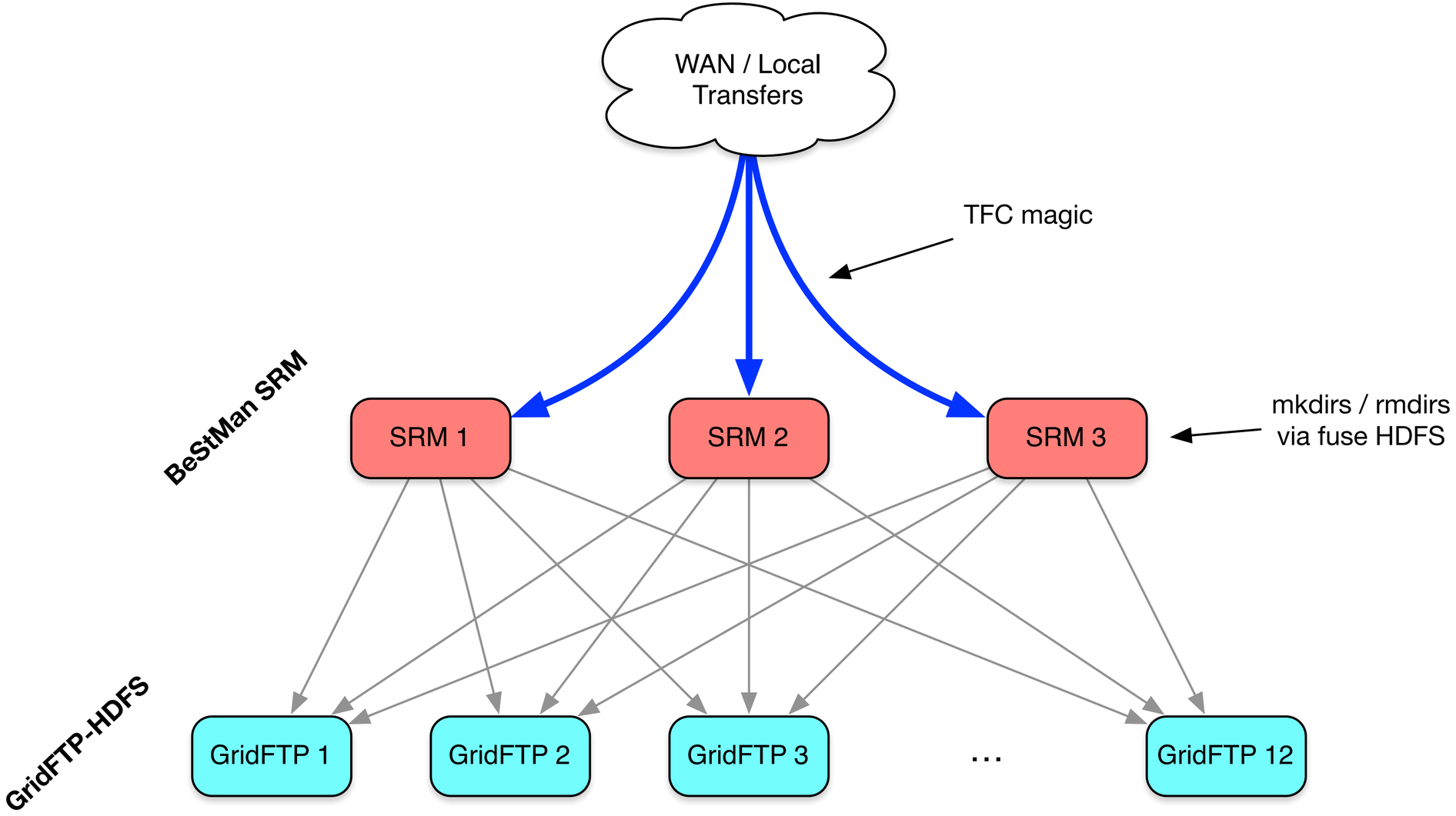
HEPiX Fall 2016

Garhan Attebury

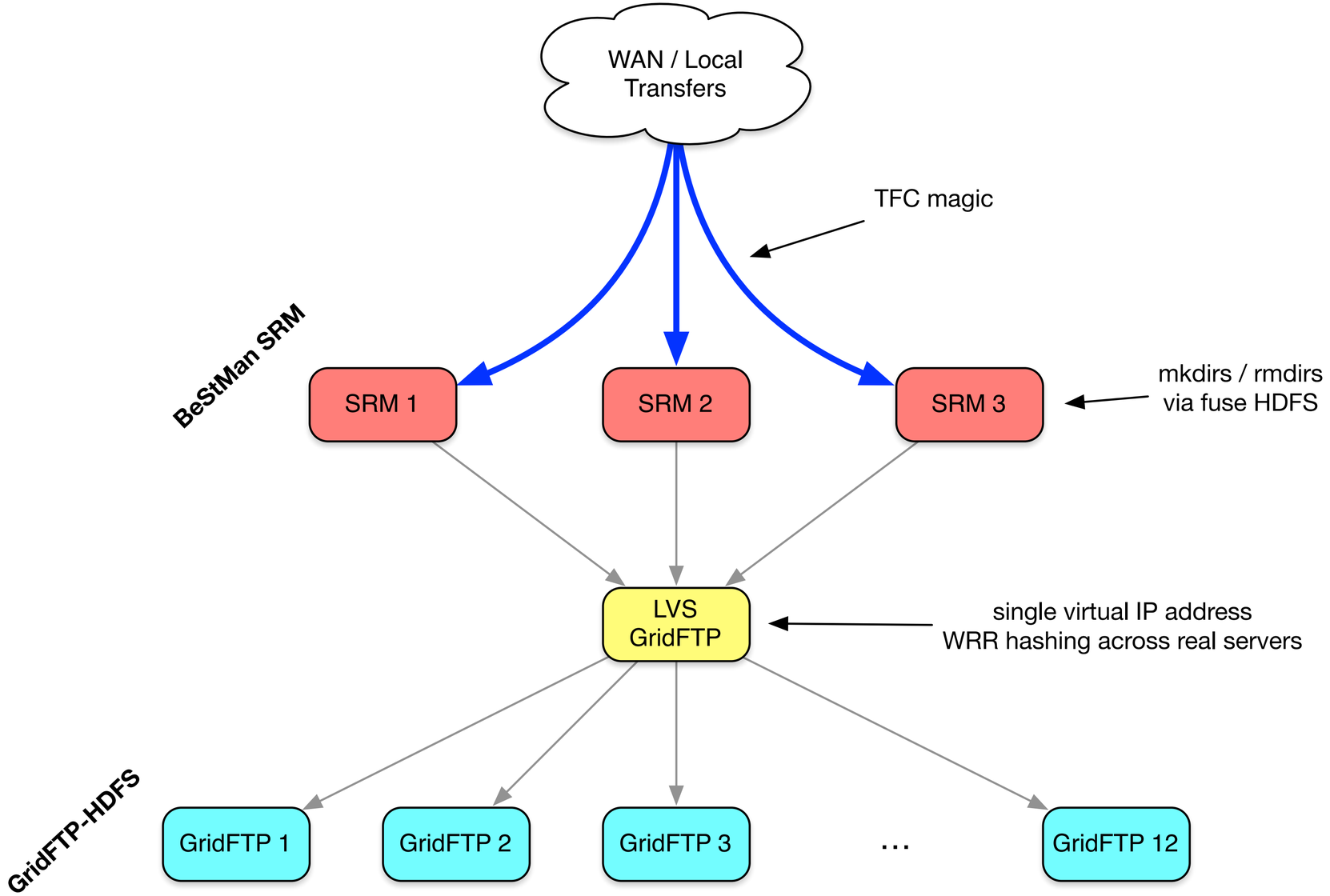
Holland Computing Center

- Provides research computing for all University of Nebraska entities
- ~28k cores in four clusters (and an openstack ‘cloud’)
- **USCMS Tier2 site**
  T2_US_Nebraska
  6,944 cores (6,304 2+GB mem slots)
  4.6PB HDFS storage
  + 1,500 slots and ~2PB storage ‘soon’
  100Gb ESnet/LHCONE/Internet2
  IPv6 wherever possible
- CMS Tier3 for UNL, KU, KSU
- Heavy involvement with the OSG

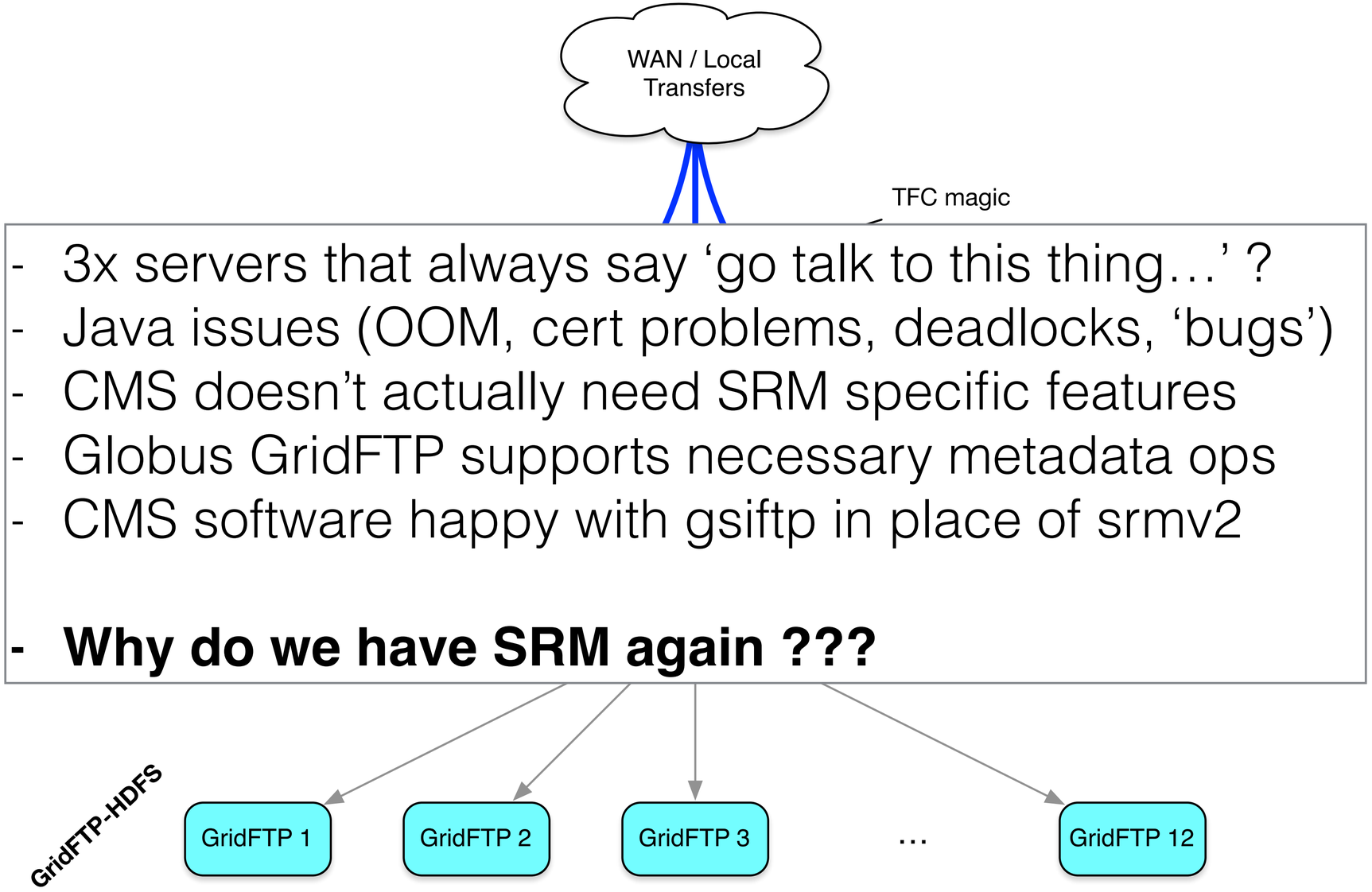
In the beginning …



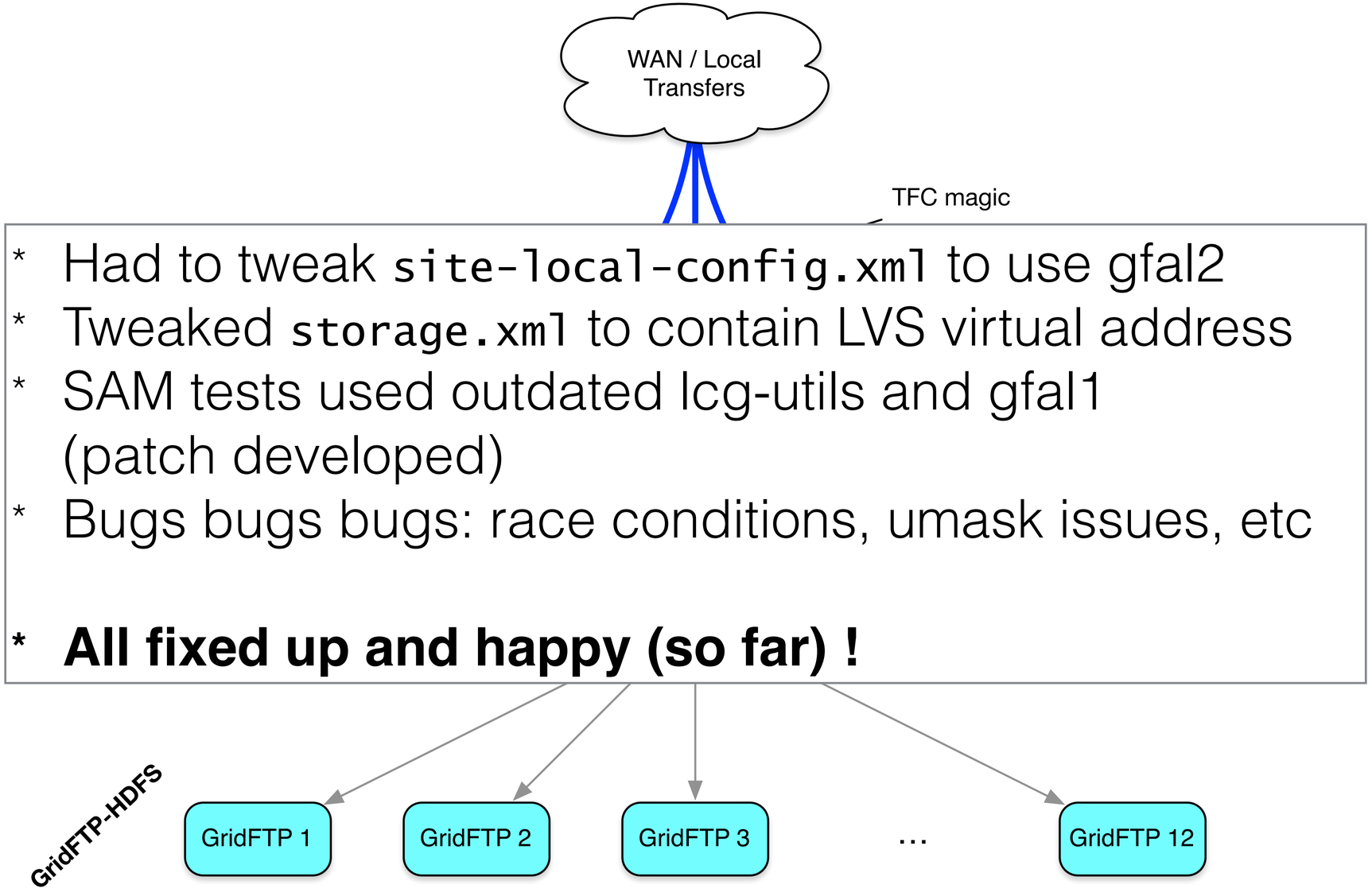
SRM + LVS balanced GridFTP



SRM + LVS balanced GridFTP



SRM + LVS balanced GridFTP

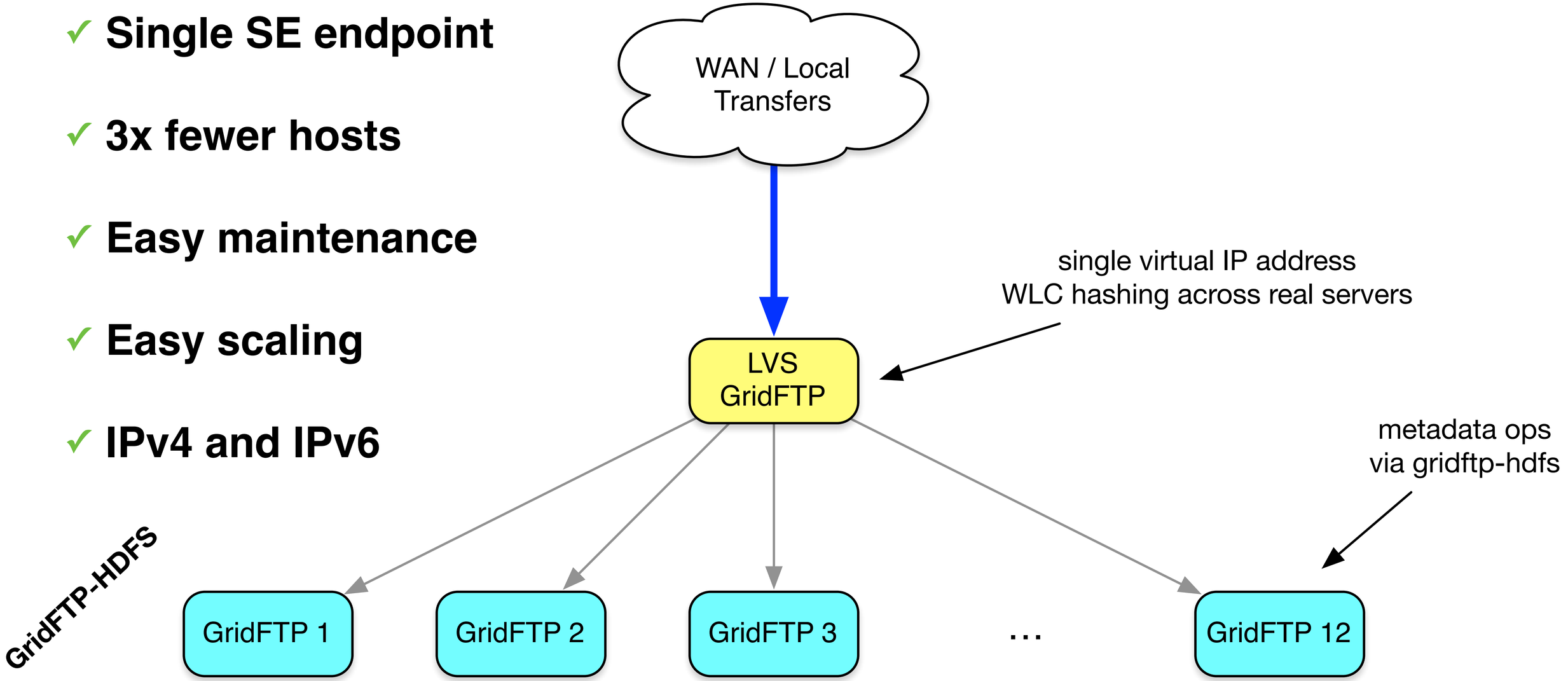


* Had to tweak `site-local-config.xml` to use gfal2
* Tweaked `storage.xml` to contain LVS virtual address
* SAM tests used outdated lcg-utils and gfal1 (patch developed)
* Bugs bugs bugs: race conditions, umask issues, etc

* **All fixed up and happy (so far) !**

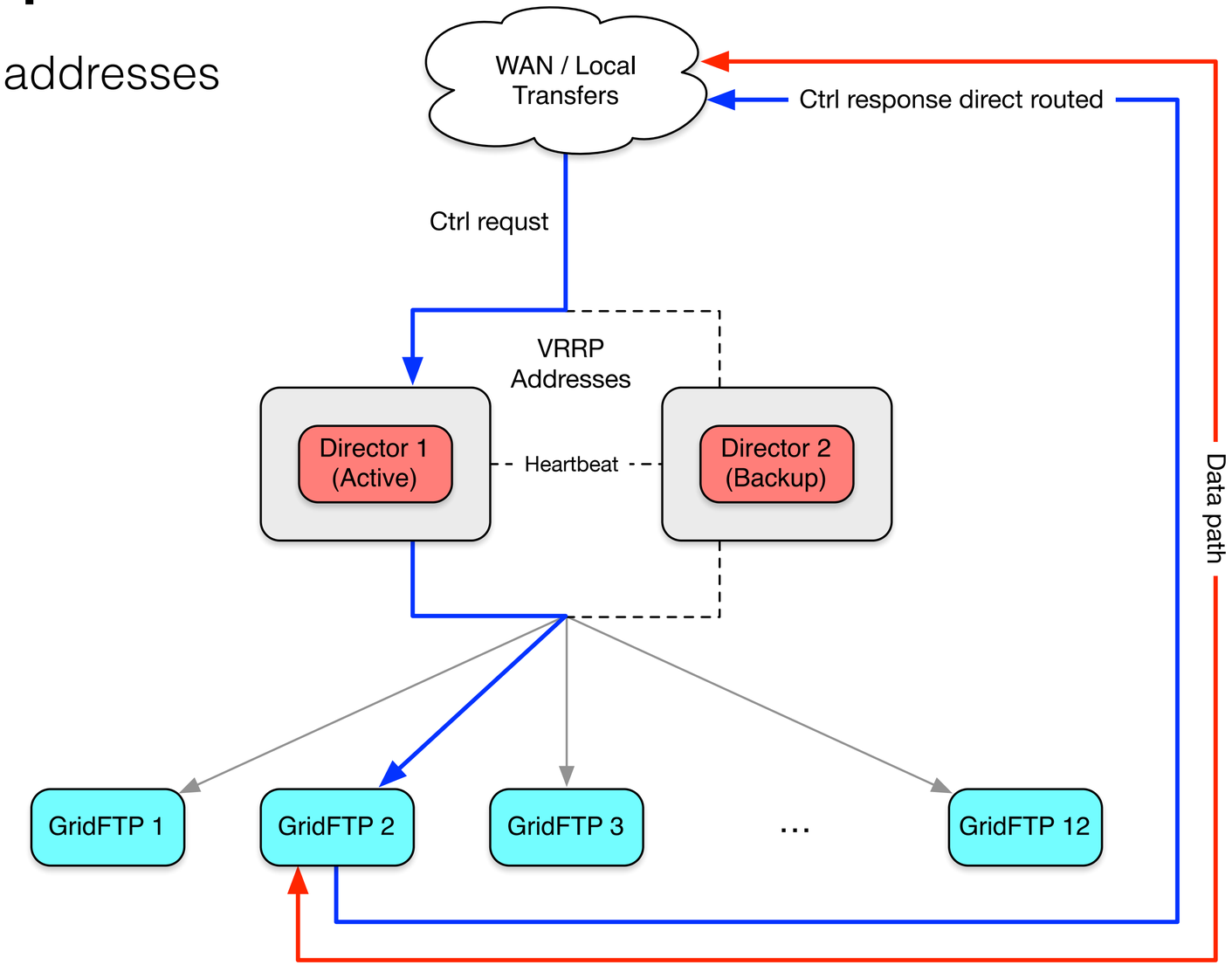
LVS GridFTP Only

- ✓ **Single SE endpoint**
- ✓ **3x fewer hosts**
- ✓ **Easy maintenance**
- ✓ **Easy scaling**
- ✓ **IPv4 and IPv6**



The recipe @ Nebraska

- Outside world only connects to VRRP addresses
- Initial response (control) direct routed
- All data path bypasses directors
- Both IPv4 and IPv6 VRRP addresses
- arptables to solve ARP problem
- Naive `rc.local` scripts to add secondary IP addresses
- Puppet configs described here



The recipe @ Nebraska

- **Documentation**:
  http://www.linuxvirtualserver.org/
  Red Hat Load Balancer Administration (EL7 version)
  https://github.com/gattebury/gridftp-with-lvs

- **Hardware side:**
  2x LVS ‘directors’ (cheap Pentium D class calculators)
  12x 10Gb GridFTP-HDFS ‘realservers’
  Flat Layer 2 network between directors/realservers

- **Configuration:**
  Keepalived via Puppet: github.com/arioch/puppet-keepalived
  Direct routing with WLC (Weighted Least Connection)

Updating base OS (never soon enough)

"Enterprise" means the package/kernel you have is
always one notch behind than the one you want

- RHEL5 —> RHEL6 transition
  - Ran SL5 chroot environment on SL6 worker nodes (thanks to built in Condor support)
  - Lacking tooling for maintaining/updating chroots
  - Required random bind mounts
    autofs + cvmfs + bind = lots of (╯°□°)╯︵┻━┻

Updating base OS (round two)

- RHEL6 —> RHEL7 transition
  - Docker is trendy, lets be trendy \o/
  - Built in support for containers… both EL7 and Condor
  - Lots (no really, A LOT) of the industry doing it already
  - … sounds like a win?

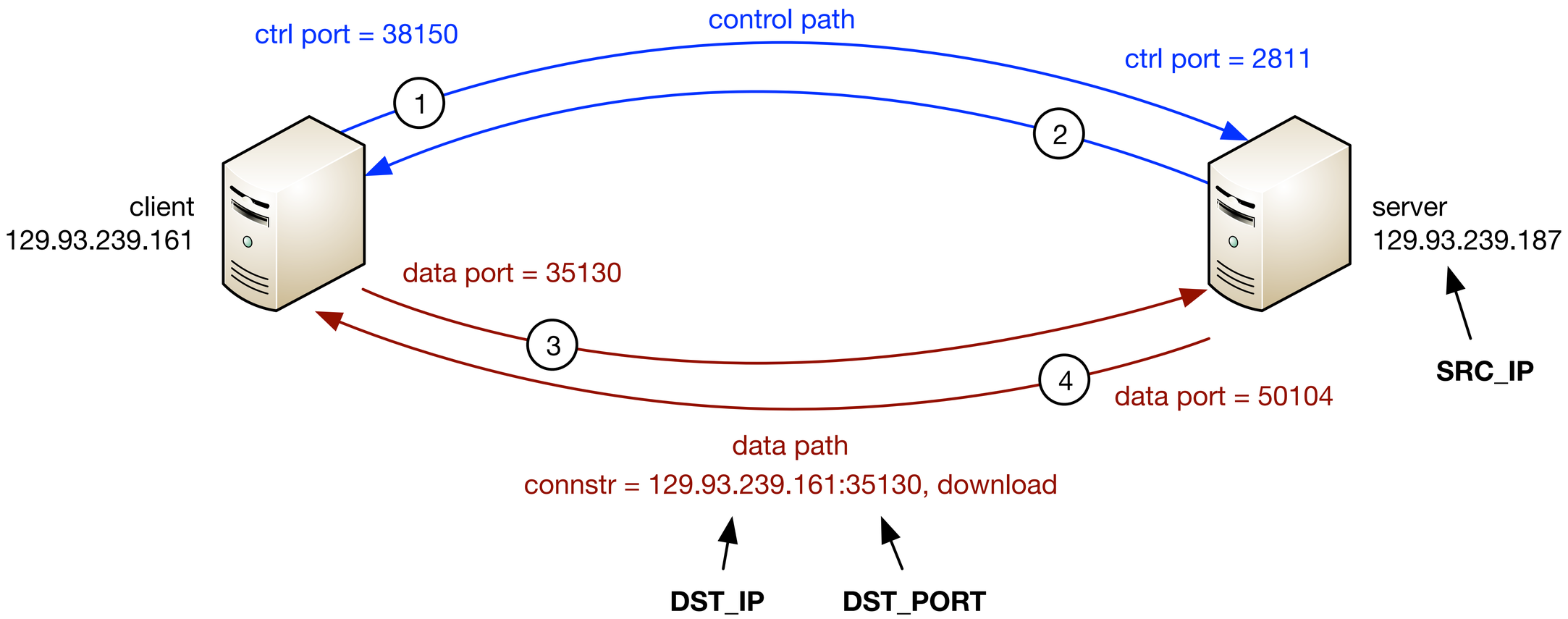
Docker + EL7 @ Nebraska

- All CentOS 7.2 - CE, SE, and workers
- Workers are fed 8-core + 20GB condor jobs,
  one container per job — Nebraska is multicore only
- Dockerfile at: https://github.com/unlhcc/docker-osg-wn-el6
  Docker hub at: https://hub.docker.com/r/unlhcc/osg-wn-el6/
- **Writeup and presentation from Derek Weitzel:**
  https://djw8605.github.io/2016/05/18/hiding-all-the-details-grid-jobs-in-docker/
  https://research.cs.wisc.edu/htcondor//HTCondorWeek2016/presentations/WedWeitzel_DockerGridJobs.pdf
- Performance penalty < 5% of native (SL6 container on CentOS 7 host)
- Flexibility of container placement (even some in Anvil)
- Happy with it for multiple months, no complaints

SDNification (?) of transfers

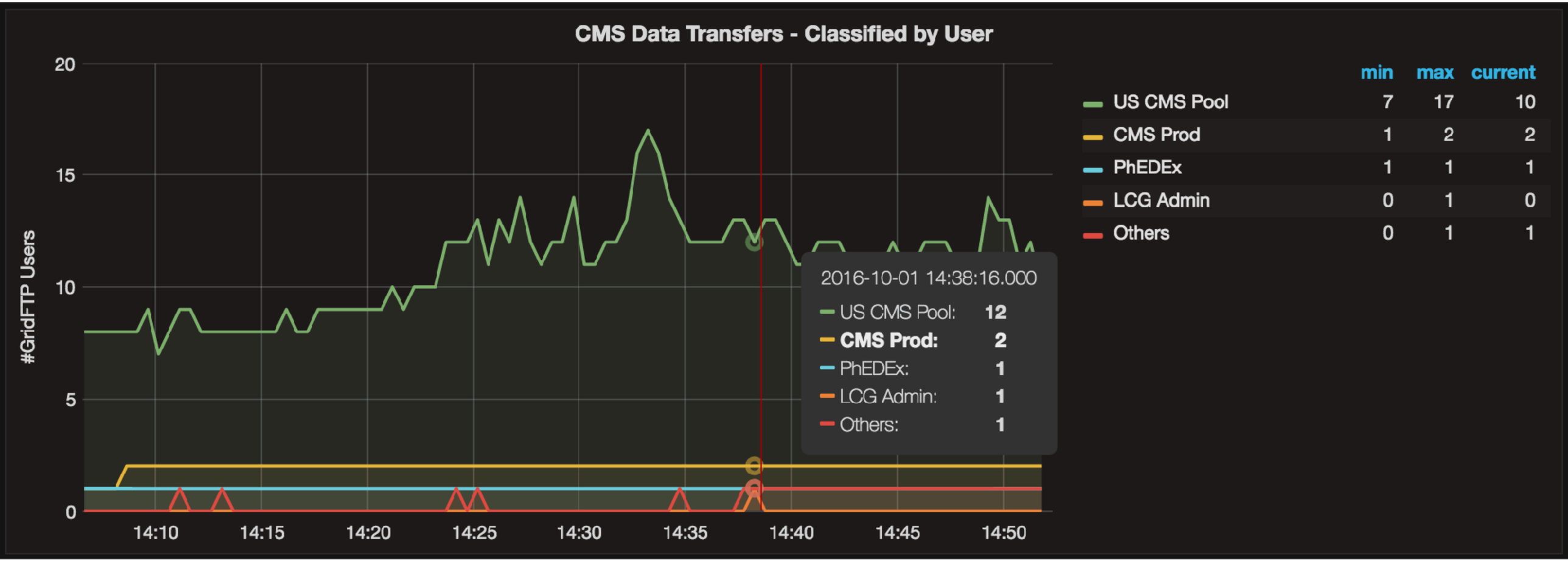
- Visibility into transfers (who? what?)
- Accounting of transfers
- Steering of transfers (automagically)

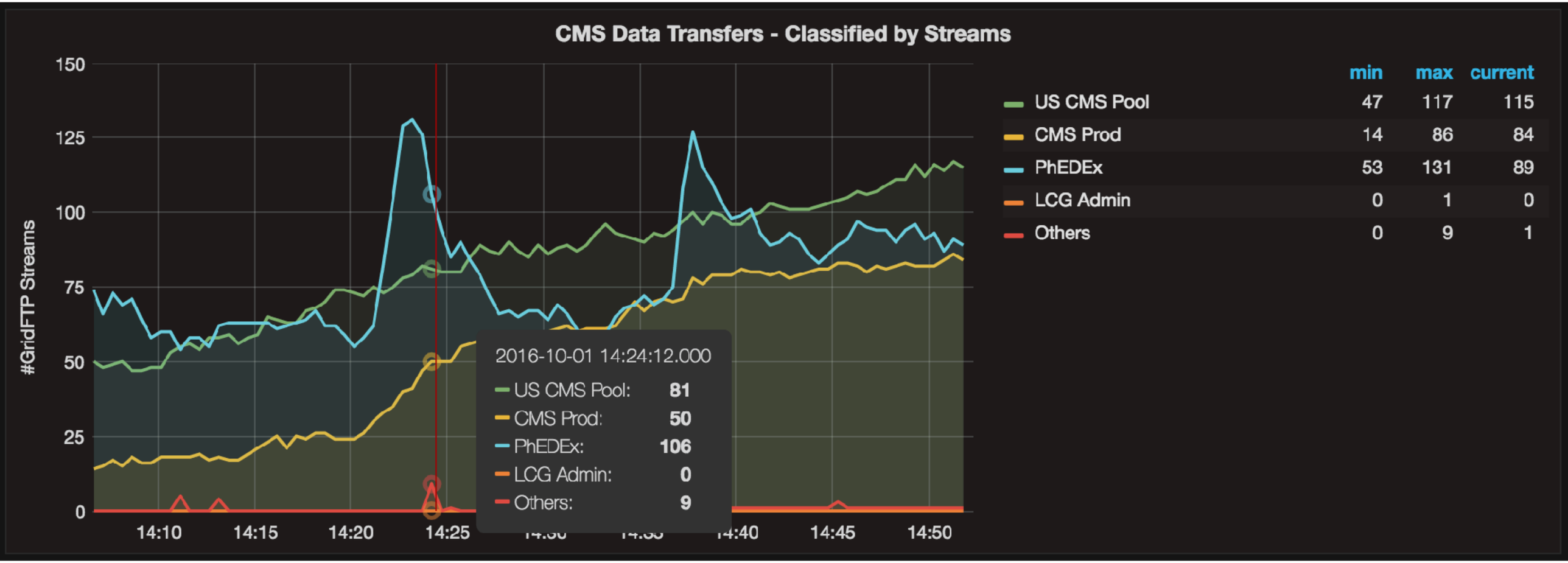
GridFTP + XIO Callout + Openflow



XIO callout gives: **SRC_IP + DST_IP + DST_PORT** of data path

Source port unknown, but can assume anything from **SRC_IP** to a given **DST_PORT** is part of the calling transfer


CMS Data Transfers - Classified by User
#GridFTP Users
20
15
10
5
0
14:10
14:15
14:20
14:25
14:30
14:35
14:40
14:45
14:50
2016-10-01 14:38:16.000
US CMS Pool: 12
CMS Prod: 2
PhEDEx: 1
LCG Admin: 1
Others: 1
min max current
US CMS Pool 7 17 10
CMS Prod 1 2 2
PhEDEx 1 1 1
LCG Admin 0 1 0
Others 0 1 1


CMS Data Transfers - Classified by Streams
#GridFTP Streams
150
125
100
75
50
25
0
14:10
14:15
14:20
14:25
14:40
14:45
14:50
2016-10-01 14:24:12.000
US CMS Pool: 81
CMS Prod: 50
PhEDEx: 106
LCG Admin: 0
Others: 9
min max current
US CMS Pool 47 117 115
CMS Prod 14 86 84
PhEDEx 53 131 89
LCG Admin 0 1 0
Others 0 9 1

SDNification: Future work

- Write a similar callout for xrootd-hdfs
- Add a UID to a transfer to group streams
- Other CC*DNI related work
  - Security — bypassing IDS, SFC, etc …
- More hardware testing (table limits, ONOS quirks, hardware quirks)
- All T2 transfers behind this, steered around Bro IDS, in production