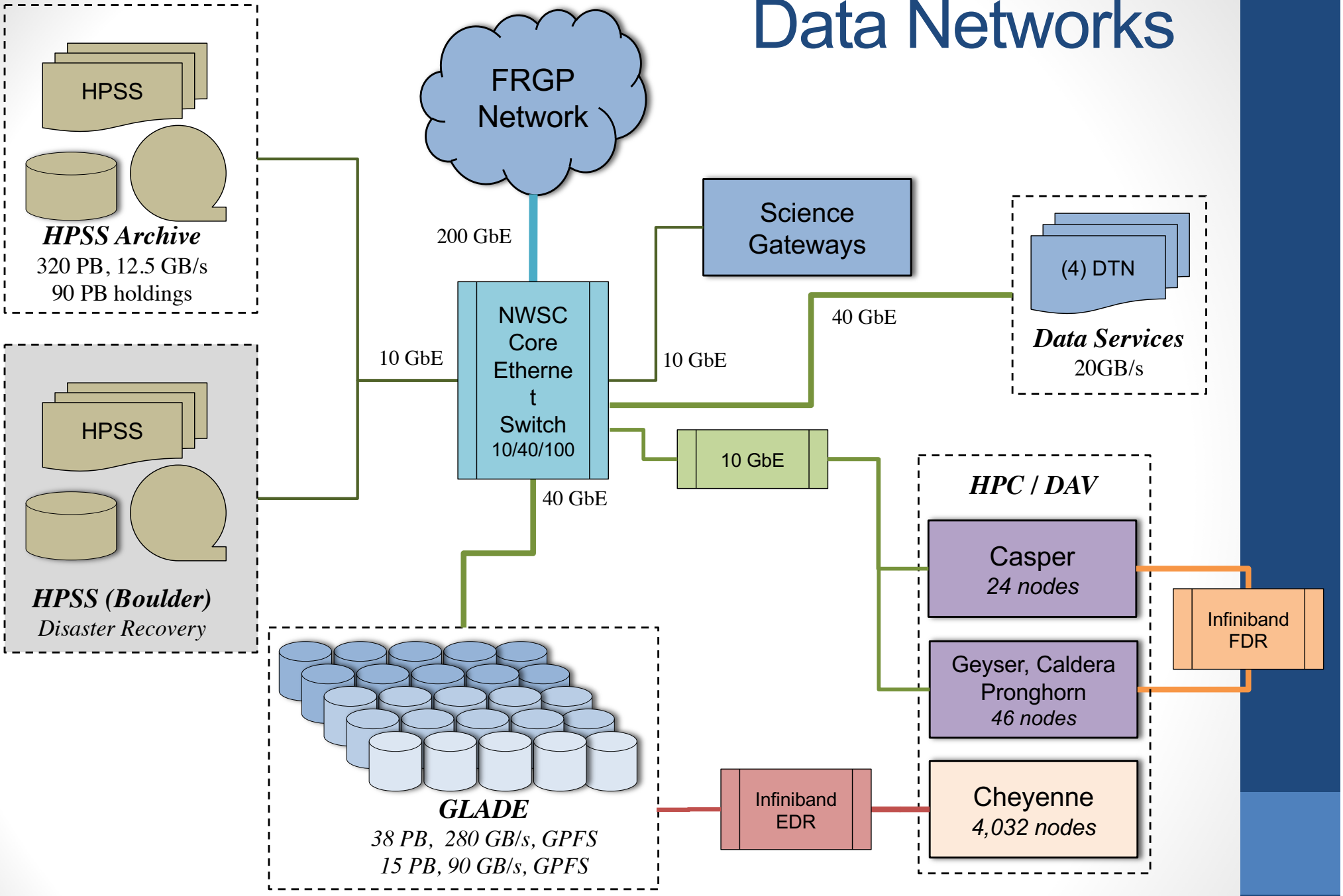
NCAR HPC Network Environment

SSUG
May 2018
Zach Mance, HPC Data Infrastructure Group

GLADE I/O Network

- *Network architecture providing global access to data storage from multiple HPC resources*
- *Flexibility provided by support of multiple connectivity options and multiple compute network topologies*
  - *10GbE, 40GbE, FDR, EDR*
  - *Full Fat Tree, Quasi Fat Tree, Hypercube*
- *Scalability allows for addition of new HPC or storage resources*
- *Agnostic with respect to vendor and file system*
- *Can support multiple solutions simultaneously*

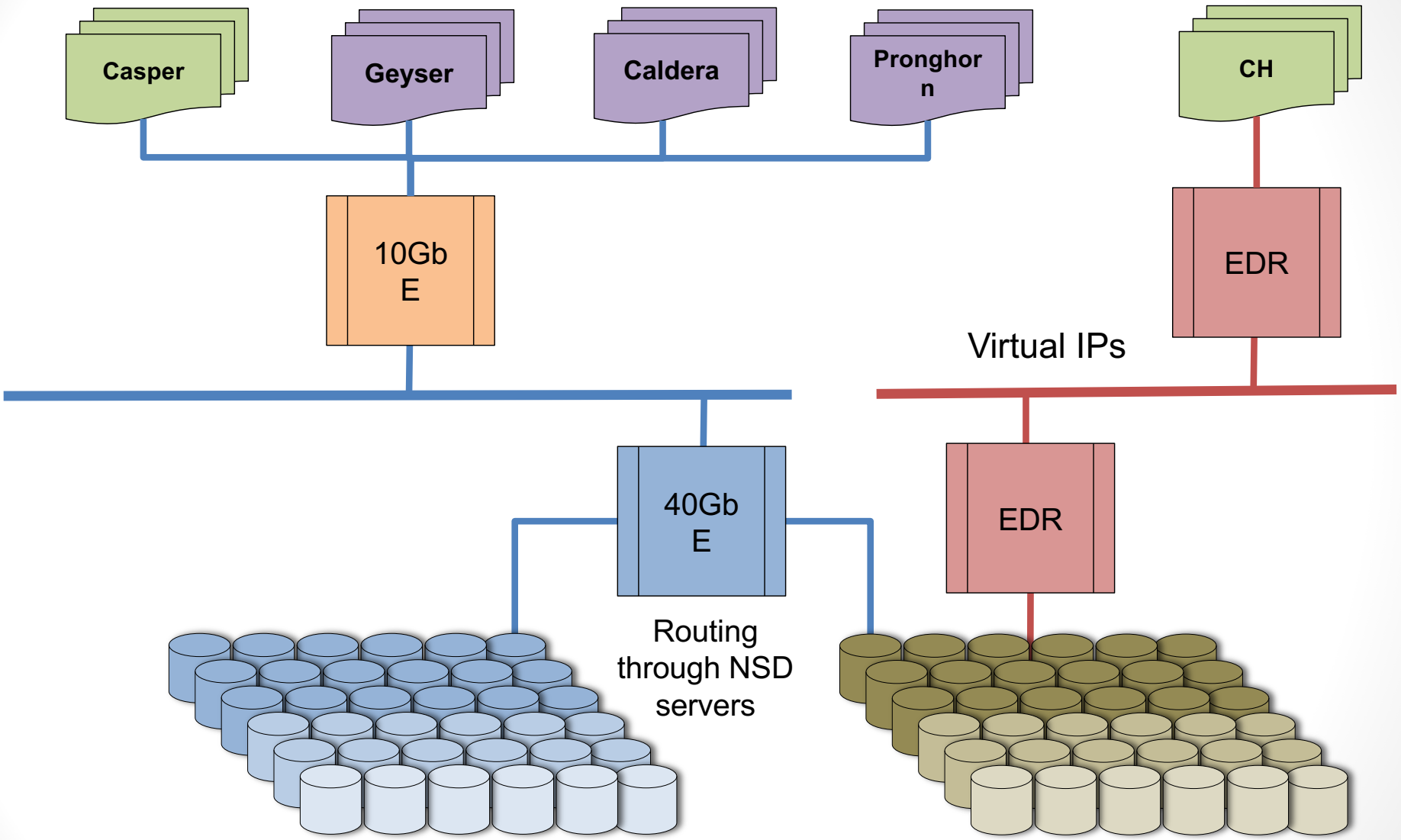
Data Networks



GLADE I/O Network Connections

- *All NSD servers are connected to 40GbE network*
- *FDR IB network*
  - *Geyser/Caldera/Pronghorn are a quasi fat tree, up/dwn routing*
  - *Capser is a fat tree, up/dwn routing*
- *DDN NSD servers are connected to the EDR IB network*
  - *Cheyenne is an enhanced hypercube*
  - *NSD VM's are nodes in the hypercube*
- *IBM NSD servers are ethernet connected only*
- *Data transfer gateways, RDA, ESG and CDP science gateways are connected to 40GbE and 10GbE networks*
- *NSD servers will route traffic over the 40GbE network to serve data to the EDR IB network*

GLADE Routing



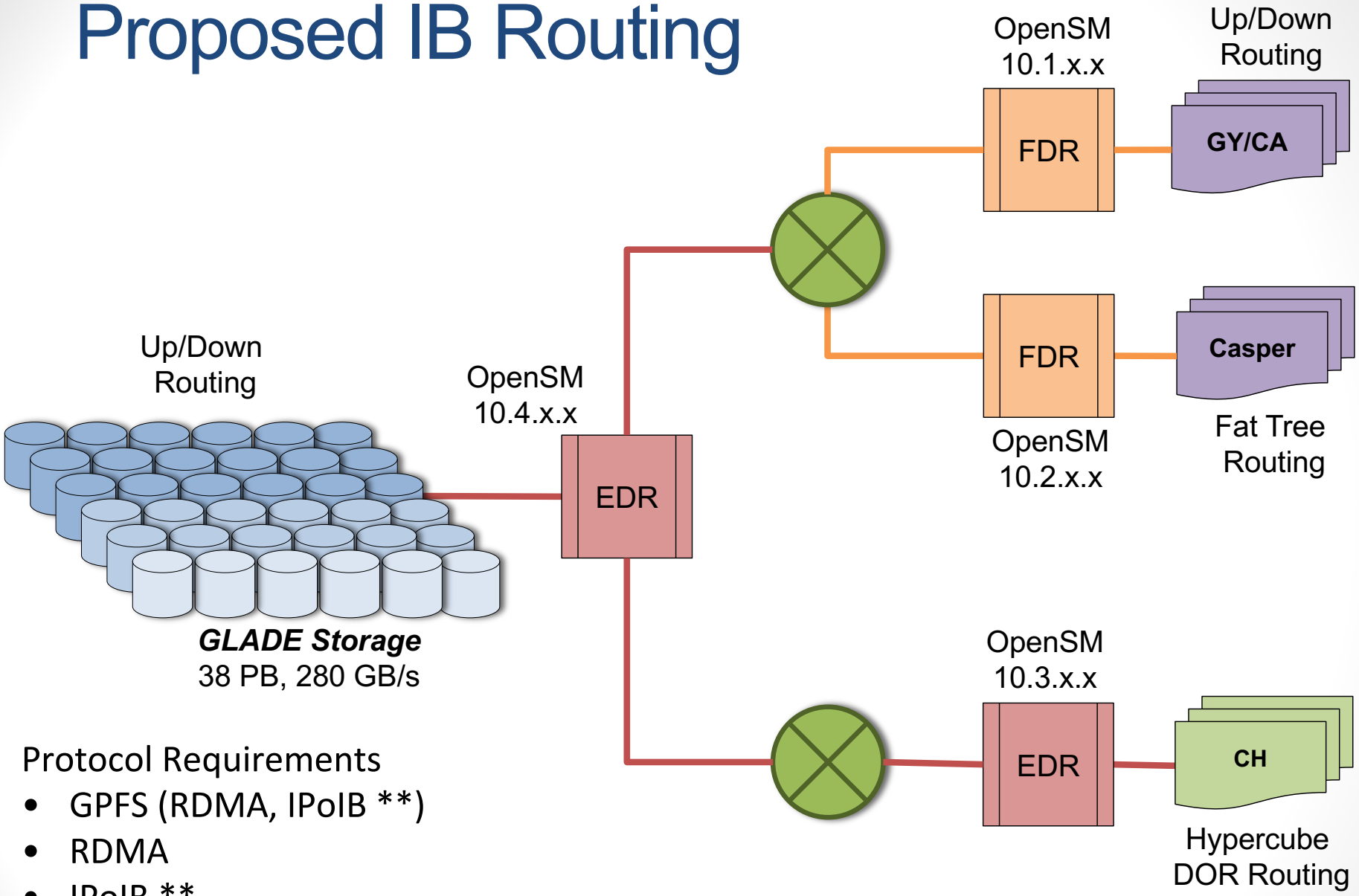
IB Network Challenges

- *Different routing algorithms between clusters*
  - *Complicated to support different routing algorithms within the same fabric manager*
  - *Complexity in troubleshooting*
- *Different generations of fabric*
  - *Concerns with degradation of performance due to complexity*
- *Storage is currently directly connected to hypercube fabric*
  - *Switch failures can bisects the fabric, storage become unreachable from some portions of compute*
  - *Outages of fabric cause loss of VERBS, sometimes have to cycle GPFS servers to recover*

IB Network Routing Pros

- *Each major cluster supports it's own IB fabric*
  - *Simplifies maintenance outages and decommissioning*
  - *Simplifies troubleshooting of fabric issues, can isolate to single fabric/routing algorithm*
  - *Can be managed by different groups*
  - *Routing algorithm is consistent within cluster*
  - *New clusters can be installed, verified, then integrated*
- *Storage remains functional through computational cluster outages*
  - *Storage cluster can block access at the router when necessary, preventing GPFS traffic to NSD's from compute*
- *Upgrades to fabric can be done per cluster*
  - *Storage fabric can remain current while compute clusters age out*

Proposed IB Routing



Protocol Requirements

- GPFS (RDMA, IPoIB **)
- RDMA
- IPoIB **
- Subnet Manager

** IPoIB – not currently supported

DataDirect NETWORKS

IBM Spectrum Scale

sgi

intel

JUNIPER NETWORKS

Mellanox TECHNOLOGIES

zmance@ucar.edu

QUESTIONS?