Fabric Computing That Works

A Case for RDMA in the WAN

Paul Grun
Chief Scientist
SystemFabricWorks
pgrun@systemfabricworks.com

11/14/2011

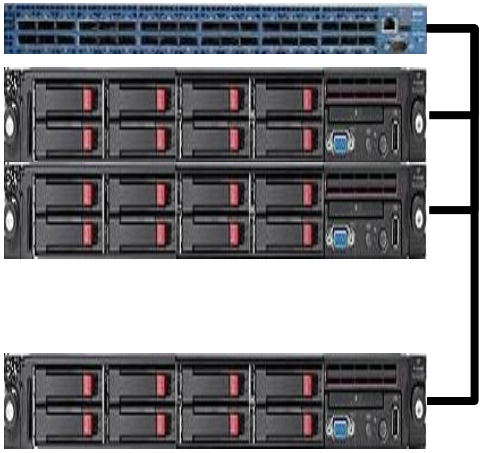
Thesis

- RDMA delivers value to HPC
- HPC has unique requirements for distributed computing
- So it makes sense to extend RDMA
- IP routers connect subnets…

  …sometimes across the WAN
- What we really want to do is to extend the subnet across the WAN

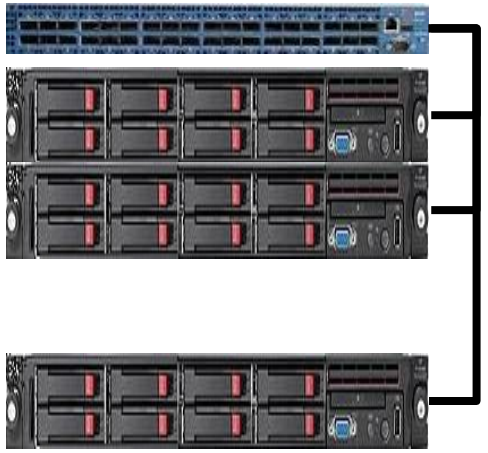
Classical HPC

- Large, scalable clusters
- Low latency interconnect for scalability
- Large storage capacity



RDMA in HPC

- Large, scalable clusters
- Low latency interconnect for scalability
- Large storage capacity

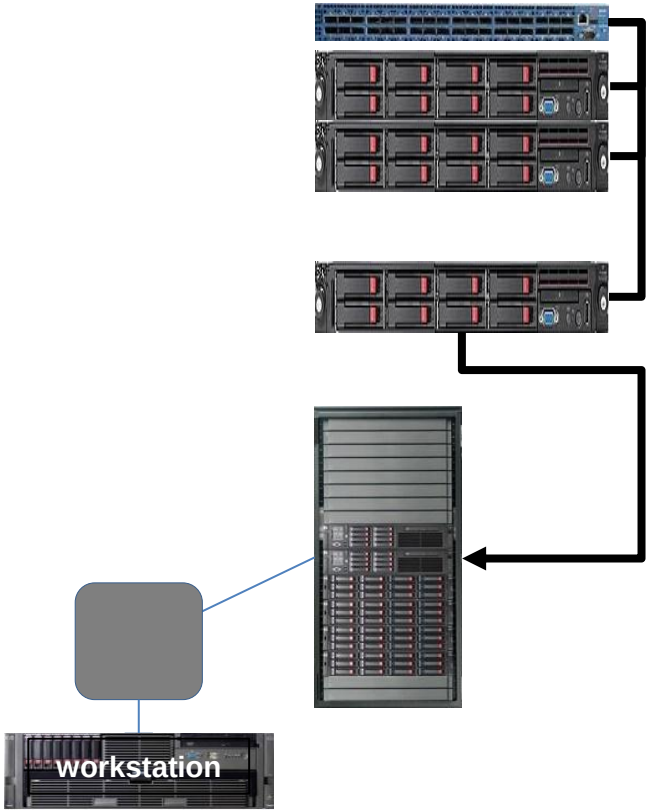


RDMA often selected for:

- Low latency (scalability)
- Bandwidth (fast storage)
- Green (better resource usage)

Using the cluster

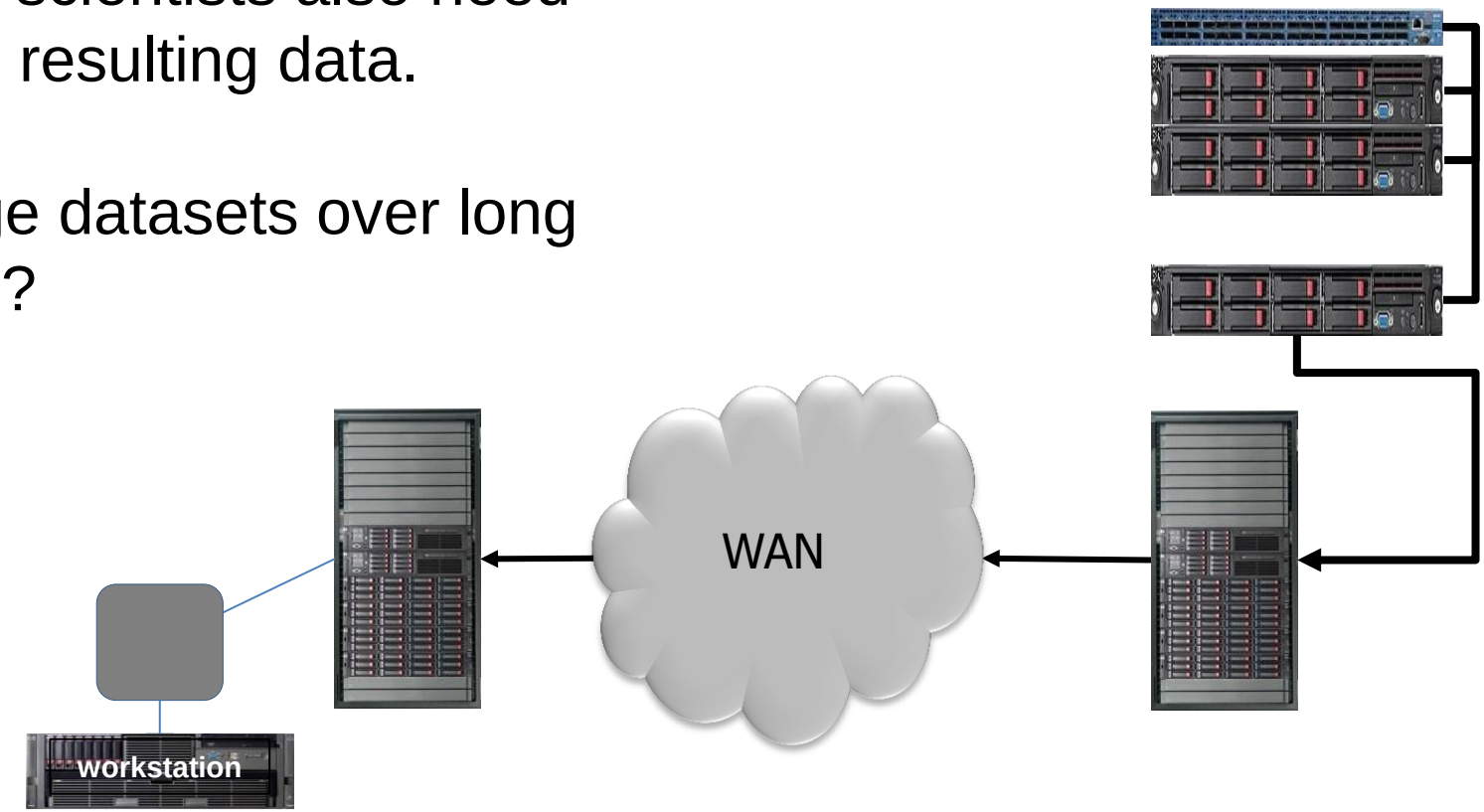
Scientists need
access to the cluster
and the resulting data



Also need to do this

(Remote) scientists also need access to resulting data.

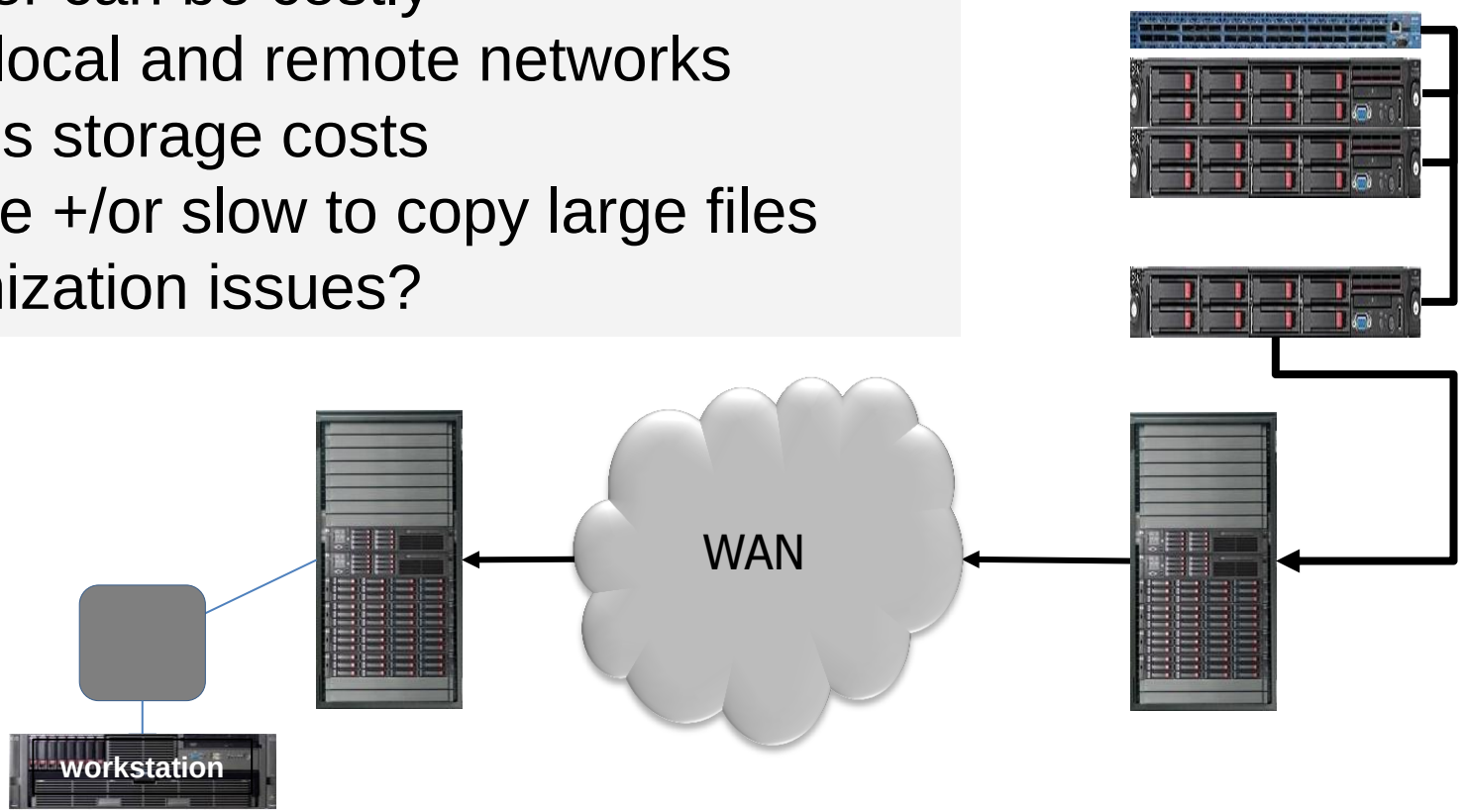
Copy large datasets over long distances?



Expensive solution

File transfer can be costly

- disrupts local and remote networks
- duplicates storage costs
- expensive +/or slow to copy large files
- synchronization issues?

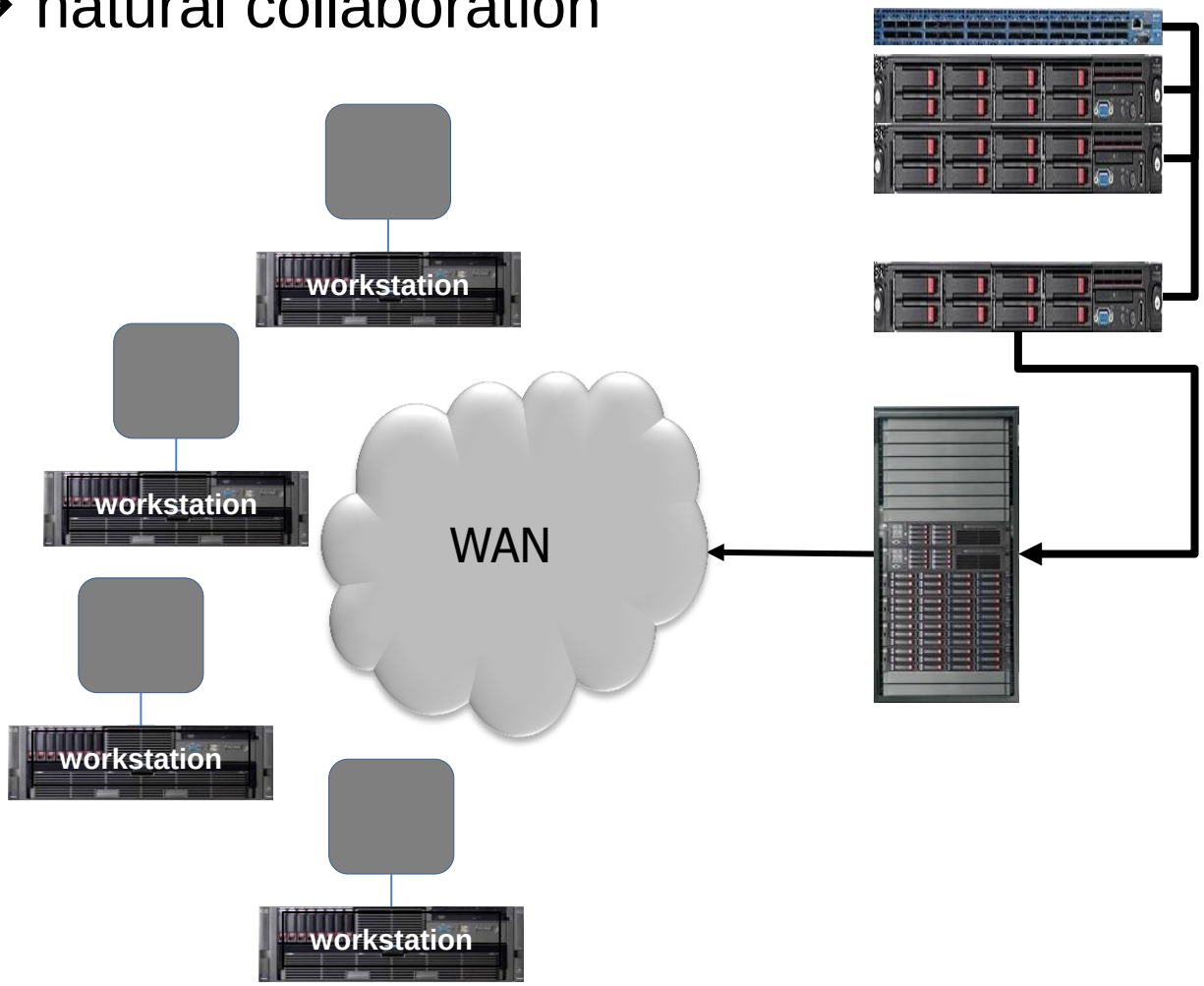


Natural Collaboration

Real time sharing → natural collaboration

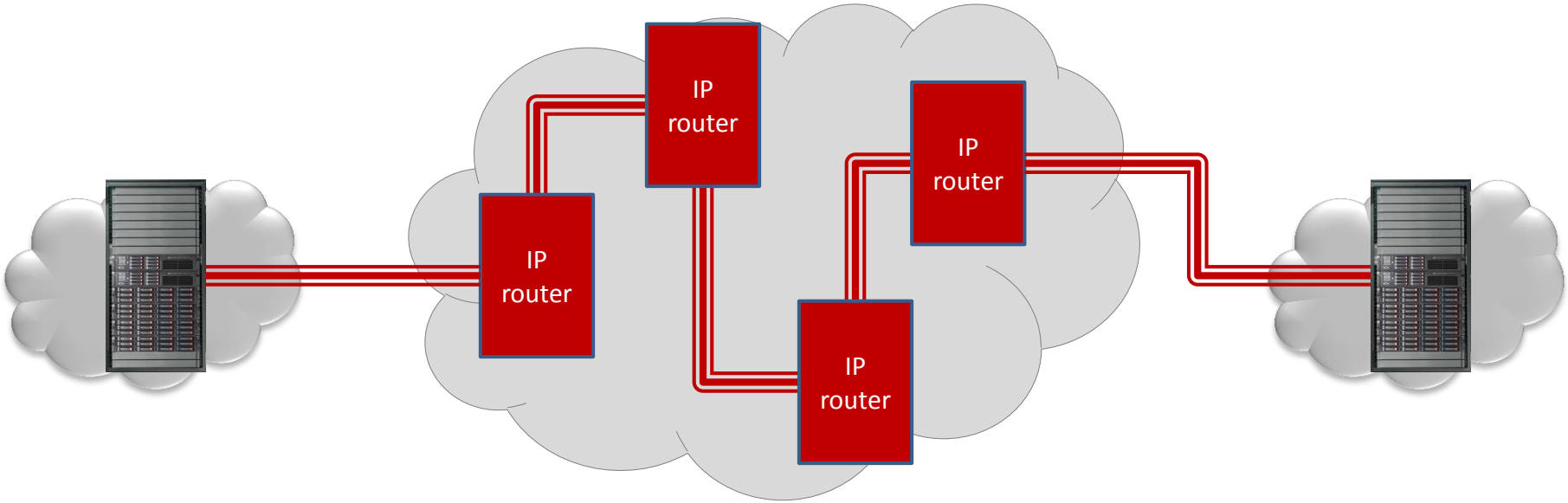
But hard to do…

…using conventional networks



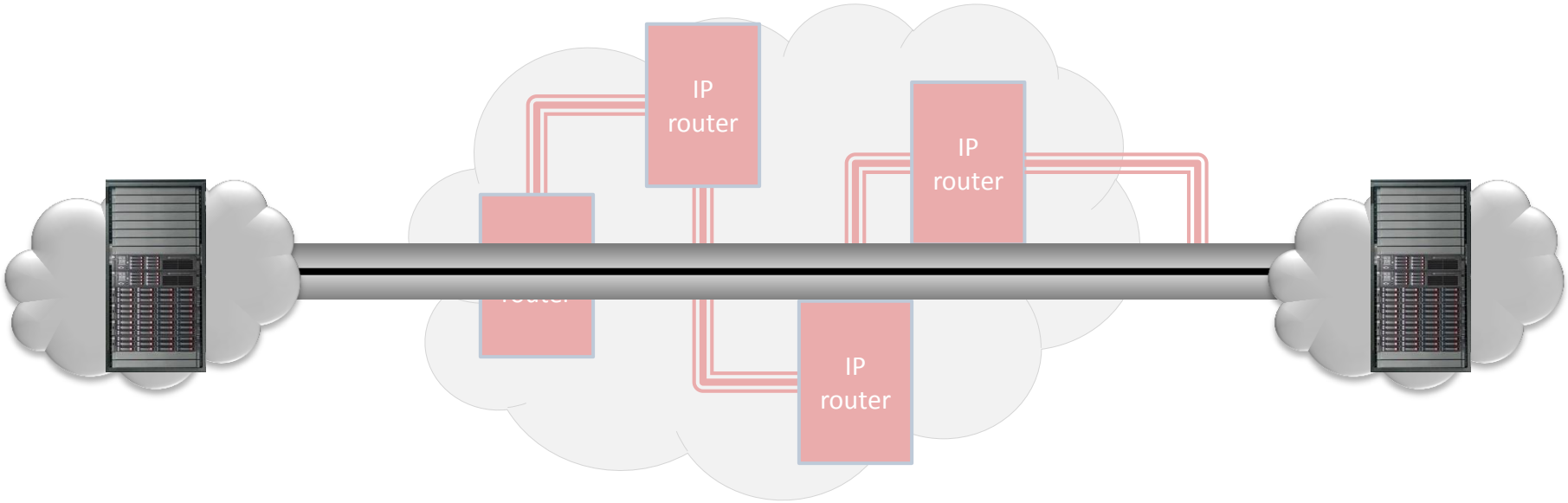
File Transfers, as a Practical Matter…

Advances in file transfer protocols are accelerating transfers



But still…
- disruptive to local and remote network traffic patterns
- duplicates storage, datasets out of sync
- consumes cycles in the endpoints for networking, consumes WAN b/w
- remote user sees delays

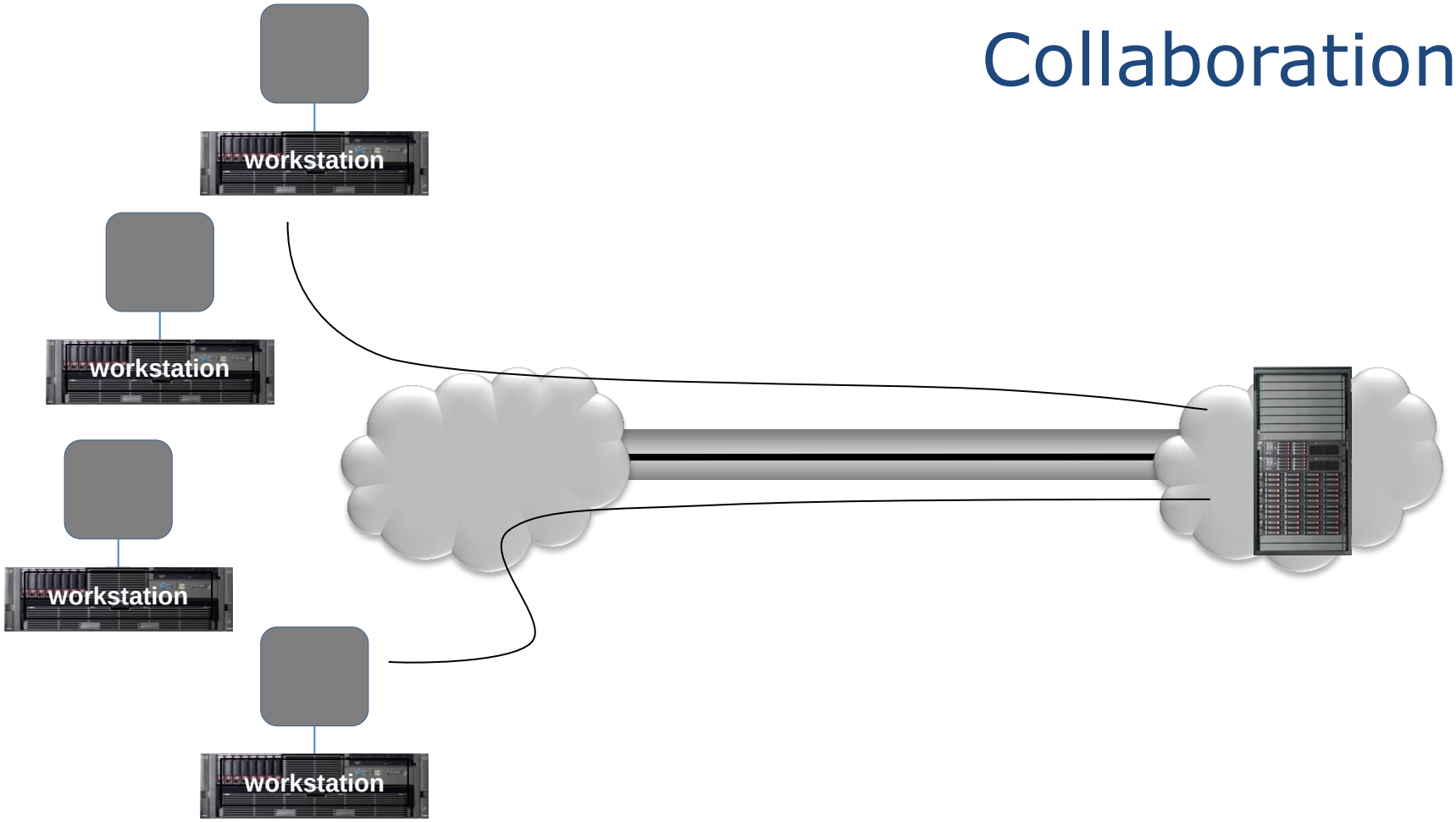
Ubiquitous connectivity…



…necessary?

L2 a better choice for limited number of endpoints?

RDMA over the WAN → Collaboration



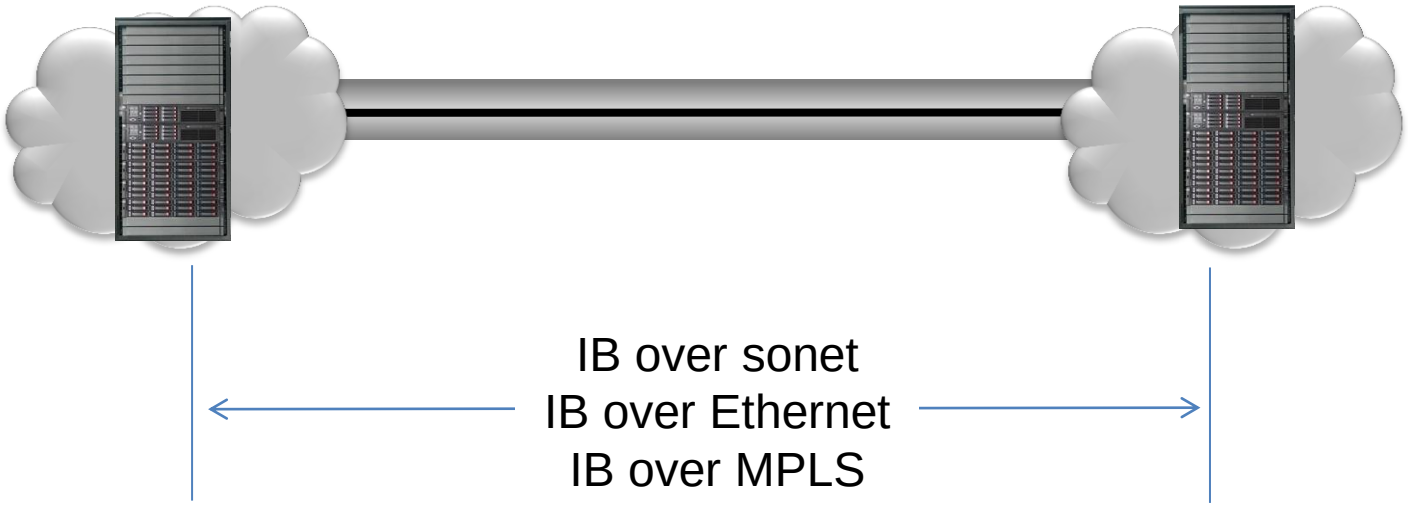
Mount the remote filesystem “as though it were local” →real time collaboration

RDMA over the WAN → Efficient File Transfer

RDMA = efficient bandwidth utilization for high bandwidth x delay product networks

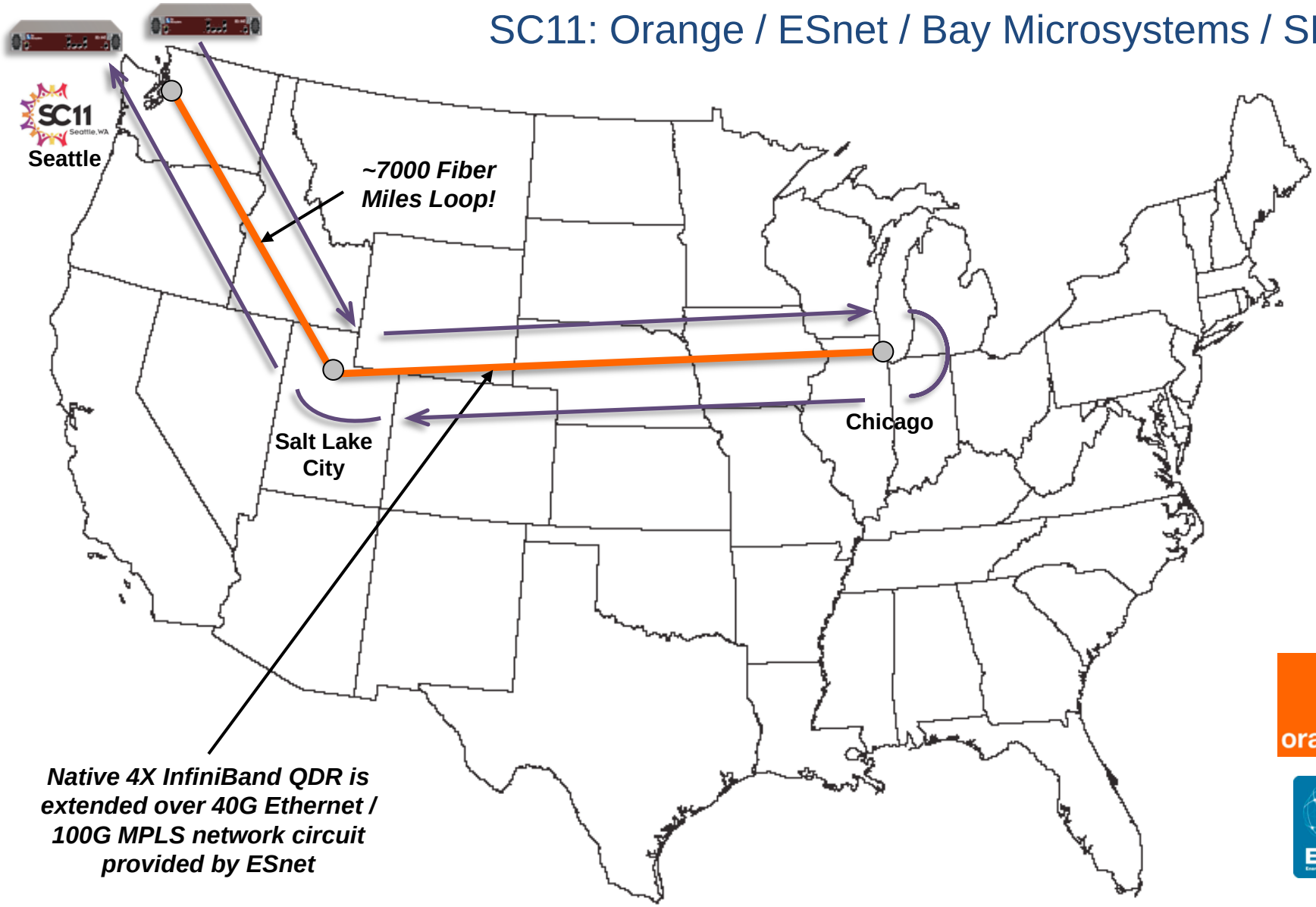
Direct data placement = low impact to local and remote subnets
Direct data placement = avoid wasted CPU cycles at either end

Works over existing WANs



Recent testing shows effective bandwidth > 96% of wire speed

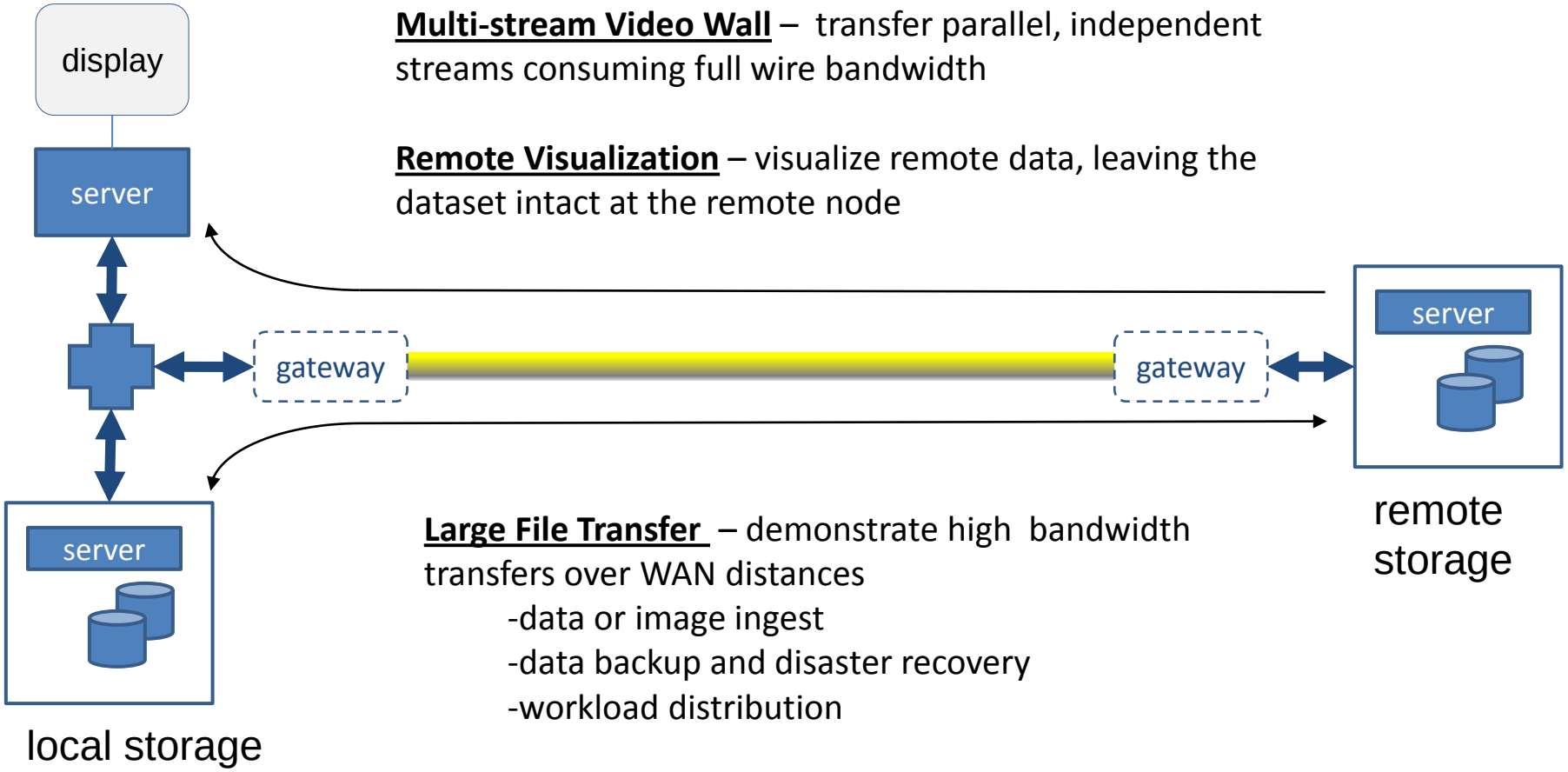
SC11: Orange / ESnet / Bay Microsystems / SFW



orange™

ESnet
Energy Sciences Network

Three Demo Workloads



Multi-stream Video Wall – transfer parallel, independent streams consuming full wire bandwidth

Remote Visualization – visualize remote data, leaving the dataset intact at the remote node

Large File Transfer – demonstrate high bandwidth transfers over WAN distances

- -data or image ingest
- -data backup and disaster recovery
- -workload distribution

- RDMA still makes sense as a cluster and storage interconnect
- Extending RDMA over the WAN gives us high performance Layer 2 connectivity
- High performance
- Enables more natural collaboration