CLEMSON
SCHOOL OF COMPUTING

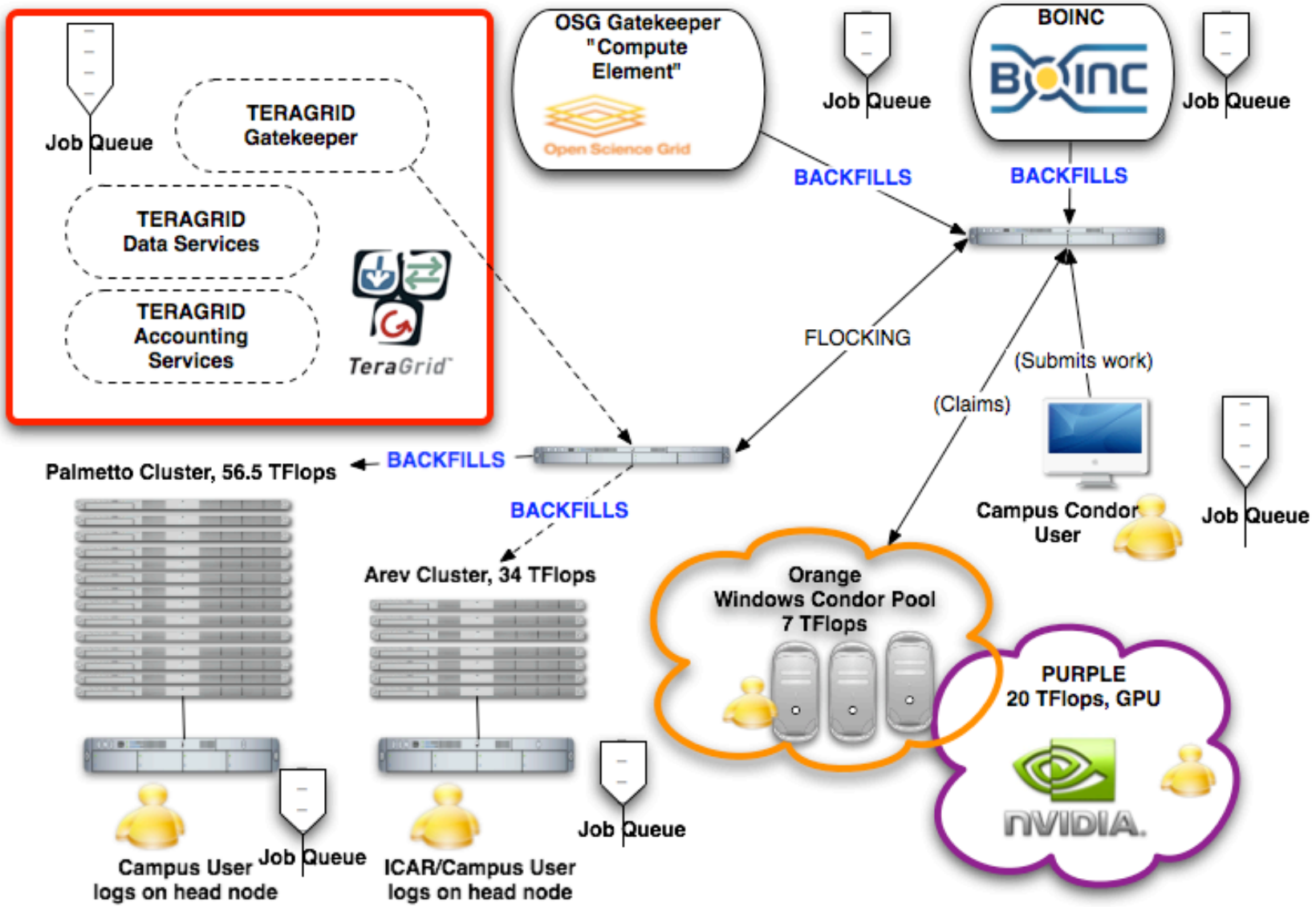
Virtual Organization Clusters

Mike Murphy, Michael Fenn and
Dr. Sebastien Goasguen, Clemson University

Open Science Grid

Clemson grid

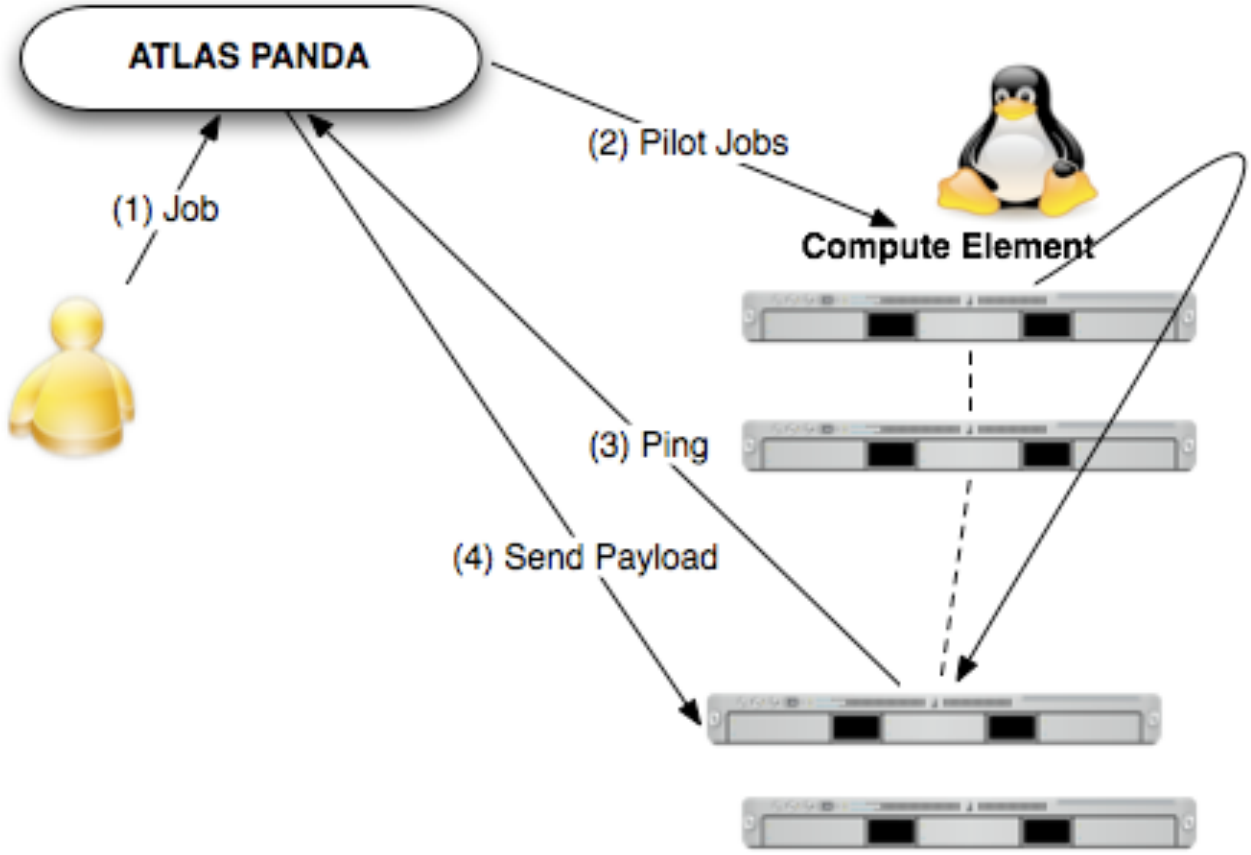
CLEMSON
SCHOOL OF COMPUTING



PANDA

CLEMSON SCHOOL OF COMPUTING

- Effort from the ATLAS VO of the LHC experiment
- Use Pilot jobs to grab slots
- Build a Scheduling Overlay over grid sites
- Maintain scheduling control especially on a per user basis.



Long Term Goal

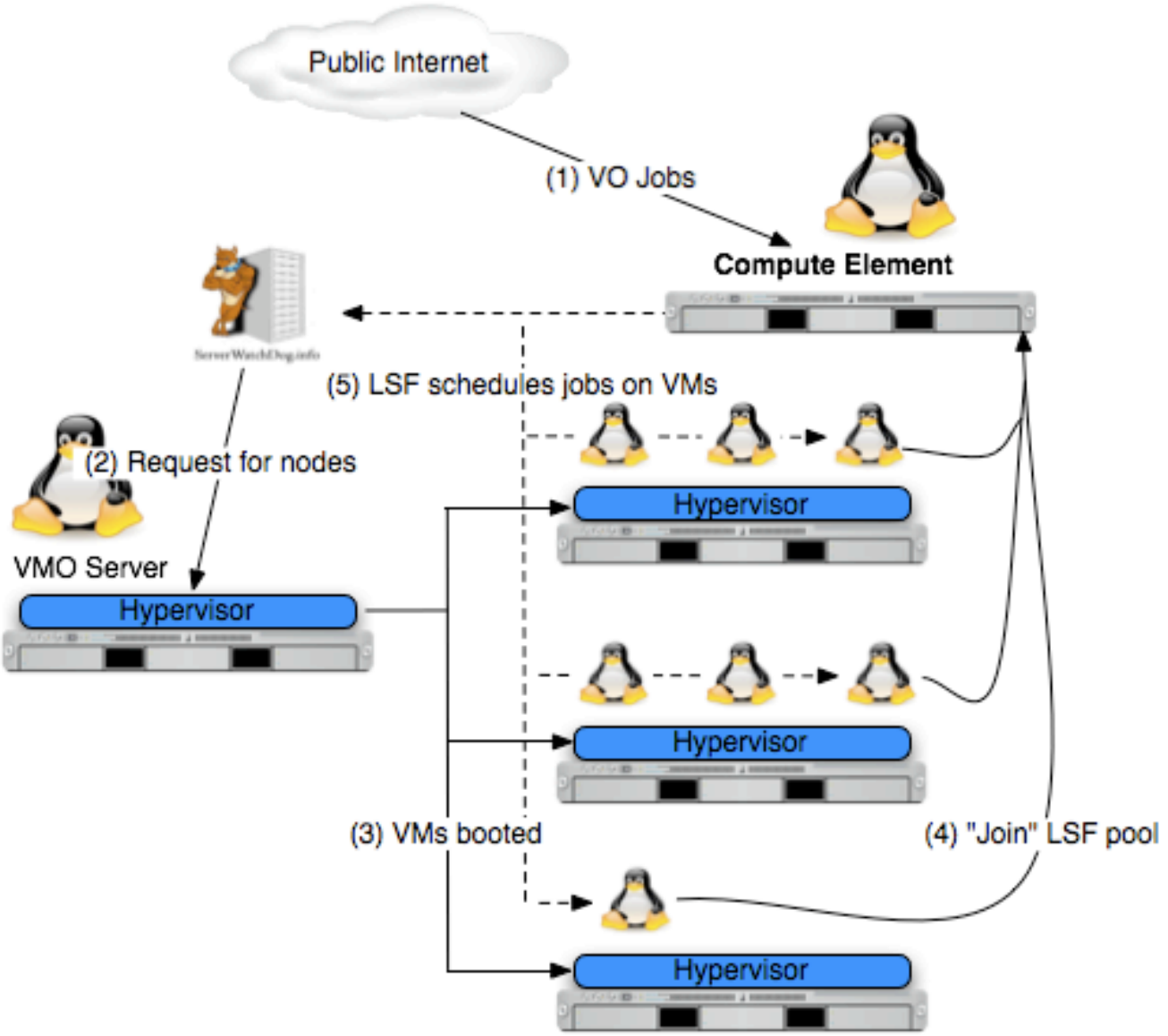
CLEMSON
SCHOOL OF COMPUTING

- **Build a VO dedicated "cluster" over WAN that is isolated from other clusters and has its own scheduling.**
- **Scheduling Overlay**
- **Nodes configured by the VO, OS is the VO's choice and all programming environment is present and up to date. There may be an extra certification step.**
- **Interoperable with existing grid mechanisms, slots are allocated via regular grid job submission techniques.**

Open Science Grid

CERN work

CLEMSON
SCHOOL OF COMPUTING

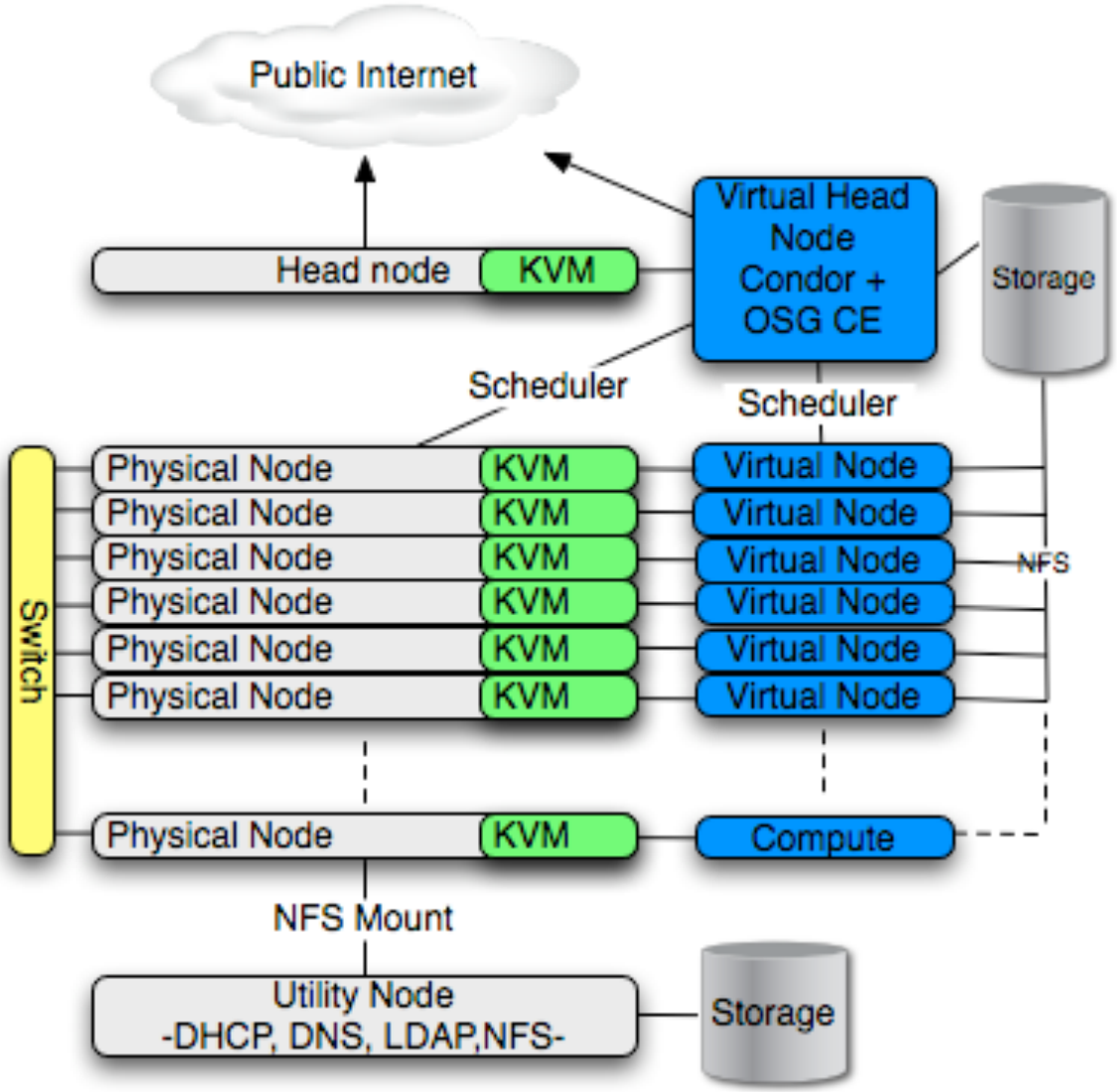


Implementation

- KVM vs. Xen for ease of use
- Normal Cluster utilities/ techniques

- NFS share
- And PVFS setup

- KVM offers a snapshot mode that gives us ability to use a single image file. Writes are temporary

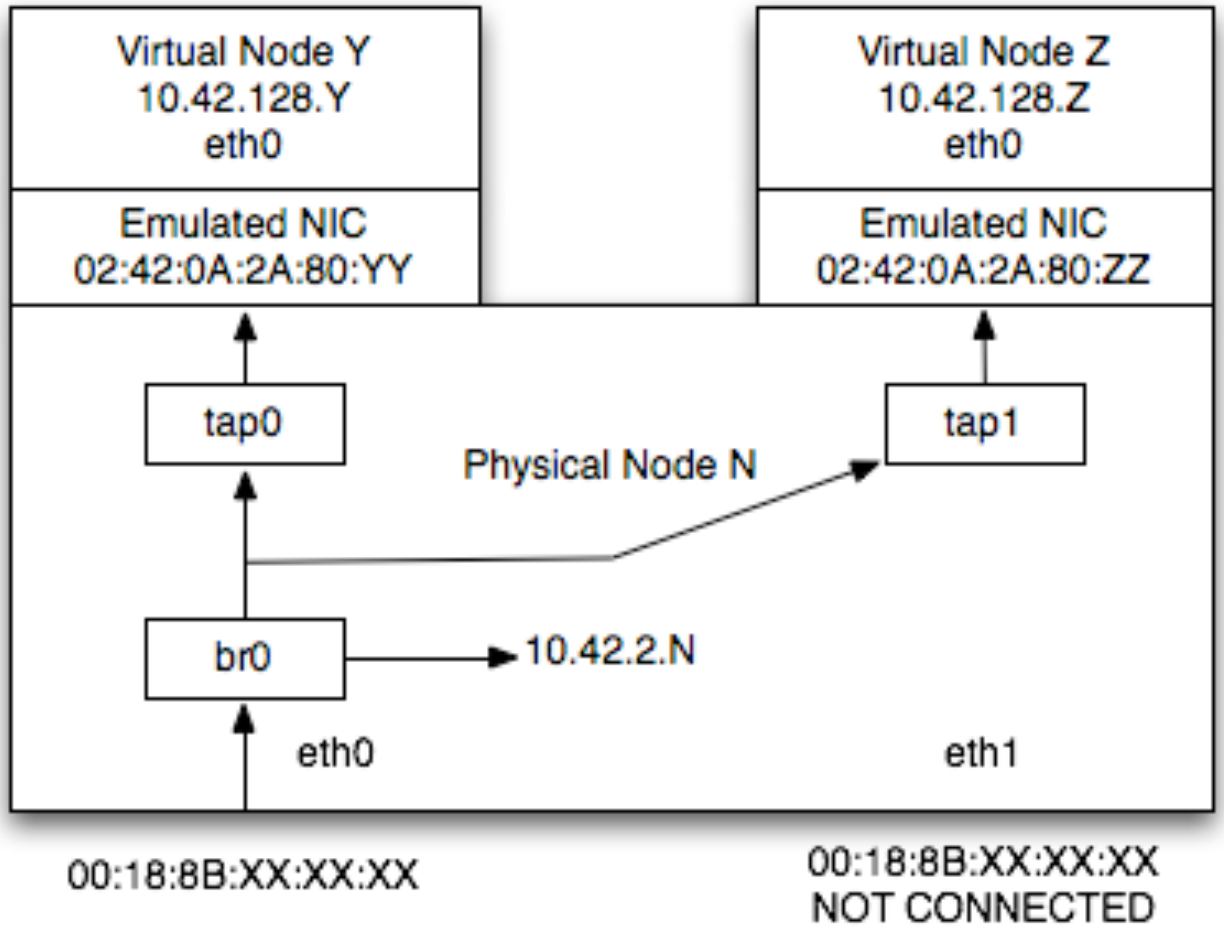


Implementation

CLEMSON SCHOOL OF COMPUTING

- Network is a challenge
- Bridge Network requires putting card in promiscuous mode
- NAT not supported by Condor

- That's why Virtual network that provide NAT traversal is good (IPOP)



Bandwidth and Latency

CLEMSON SCHOOL OF COMPUTING

- **Bridge networking and offloading network operation to CPU increases latency and reduces bandwidth**
- **Plus context switches in VM I/O**

Table 4. Bandwidth and Ping Latency

Condition	No-Load			Under Load		
Parameter	P	SB	BTB	P	SB	BTB
Iperf ($Mb \cdot s^{-1}$)	941	499	708	882	206	636
RRB ($Mb \cdot s^{-1}$)	N/A			544	24	32
Ping RTT (μs)	106	215	312	191	360	484
RRL (μs)	N/A			54	379	233

Key: P – Physical, SB – Virtual links across the Same Bridge, BTB – Virtual links from one bridge (physical host) to another, RRB – RandomRing Bandwidth, RRL – RandomRing Latency

Performance issues

CLEMSON SCHOOL OF COMPUTING

- **Virtualization introduces overhead.**
- **Is it that bad, that using VMs in grids does not make sense**

- **->Benchmark applications in various hypervisor**
- **HPL is used for CPU intensive runs**
- **HPL also provide message passing tests (MPI)**
- **HPCC provides other benchmark tests for I/O**

Open Science Grid

Performance results

CLEMSON SCHOOL OF COMPUTING

- **High Performance Linpack (HPL, 1x1, N=10300)**
- **Xen ~6.5%, KVM ~8.7%**

Table 2: Physical vs. Virtualized, Single Process

Process Grid	1x1 Physical	1x1 Xen VOC	Xen Overhead	1x1 KVM VOC	KVM Overhead
Problem Size	10300	10300	0%	10300	0%
G-HPL (GFLOPS)	7.913	7.393	6.566%	7.218	8.771%
G-PTRANS (GB/s)	0.729	0.588	19.415%	0.635	12.946%
G-Random Access (GUP/s)	0.002	0.001	35.519%	0.002	15.818%
G-FFTE (GFLOPS)	0.799	0.658	17.733%	0.461	42.370%
EP-STREAM Sys (GB/s)	3.866	3.375	12.704%	3.808	1.491%
EP-STREAM Triad (GB/s)	3.866	3.375	12.704%	3.808	1.491%
EP-DGEMM (GFLOPS)	8.348	7.689	7.892%	7.682	7.977%
RandomRing Bandwidth (GB/s)	N/A	N/A	N/A	N/A	N/A
RandomRing Latency (μs)	N/A	N/A	N/A	N/A	N/A

Performance results

CLEMSON SCHOOL OF COMPUTING

- **High Performance Linpack (HPL, 7x4, N=58600)**
- **Xen ~23%, KVM ~52%**

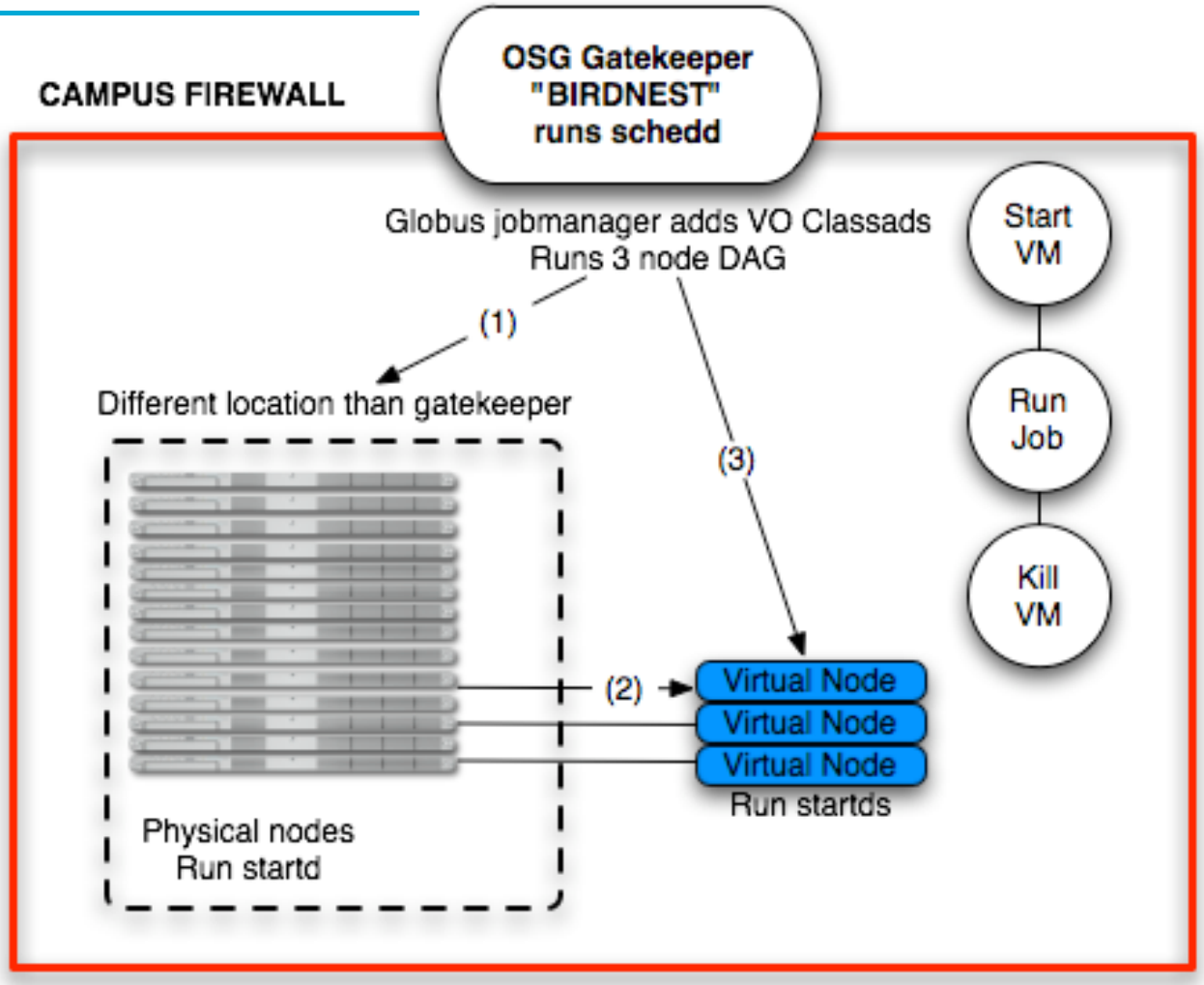
- **Latency Kills KVM.**

Table 4: Physical vs. VOC, Two VMs per Physical Node (32 processes)

Process Grid	7x4 Physical	7x4 Xen VOC	Xen Overhead	7x4 KVM VOC	KVM Overhead
Problem Size	58600	58600	0%	58600	0%
G-HPL (GFLOPS)	169.807	130.862	22.935%	81.401	52.063%
G-PTRANS (GB/s)	0.867	0.830	4.302%	0.447	44.968%
G-Random Access (GUP/s)	0.014	0.011	22.941%	0.004	70.643%
G-FFTE (GFLOPS)	2.287	0.746	67.380%	1.751	23.449%
EP-STREAM Sys (GB/s)	59.046	62.382	-5.650%	73.110	-23.818%
EP-STREAM Triad (GB/s)	1.845	1.949	-5.650%	2.285	-23.818%
EP-DGEMM (GFLOPS)	8.271	7.726	6.588%	7.114	13.979%
RandomRing Bandwidth (GB/s)	0.023	0.007	68.779%	0.027	-17.148%
RandomRing Latency (μs)	74.444	125.258	67.259%	228.383	206.787%

Update: Condor and Clouds

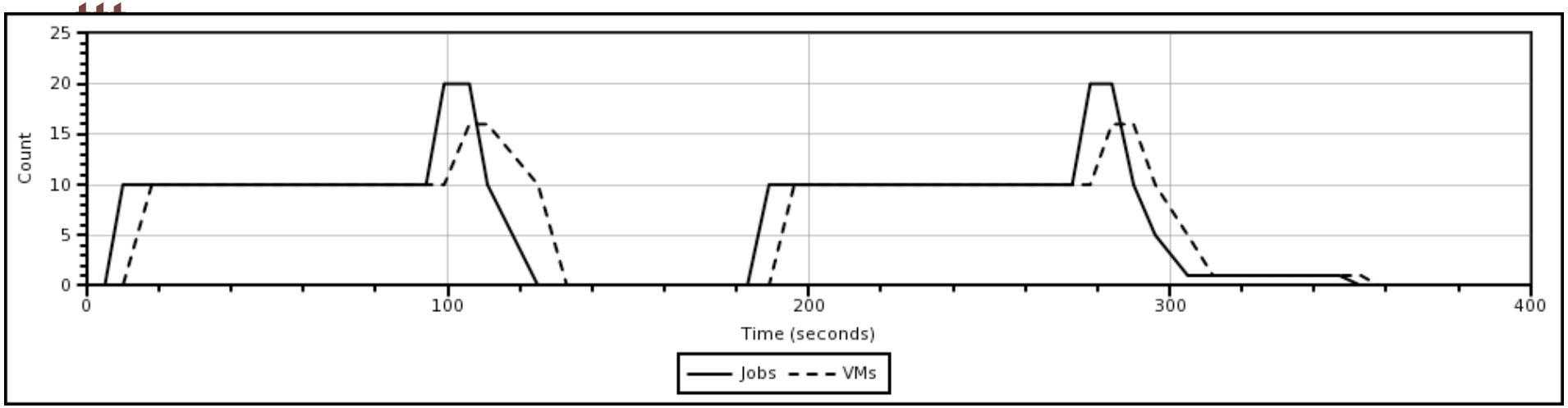
- VM as a job
- Job “glides” in VM
- VM destroyed
- “VPN” for all VMs
- Different OS/sw for each VO
- Use EC2…
- Use VM universe…

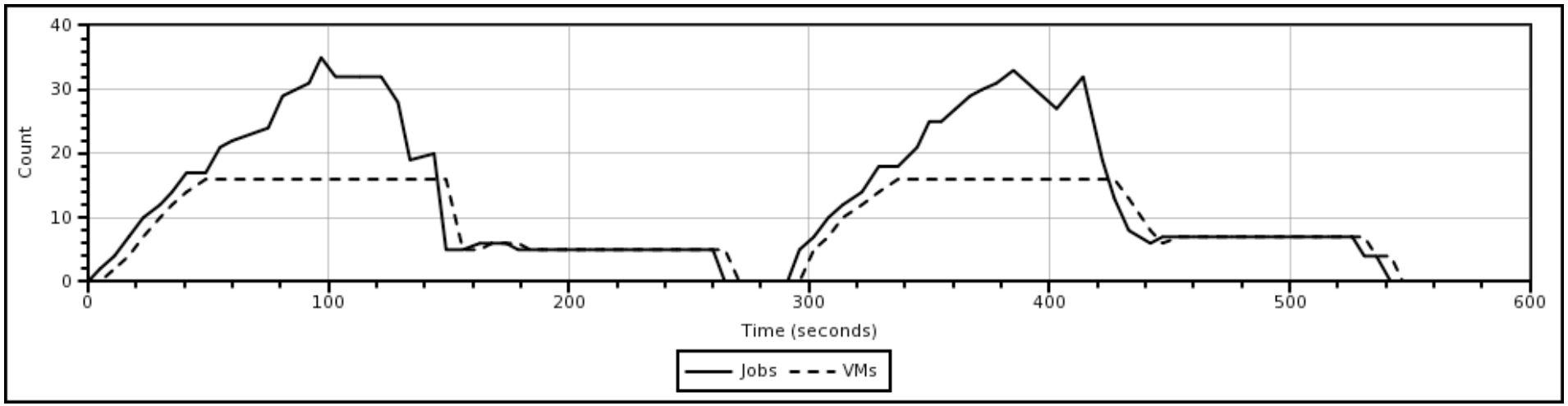


- **Use IPOP (http://www.grid-appliance.org/) to build WAN “VPN” that traverse NATs. Ability to isolate clouds in different address space.**

Update: Latest Results

CLEMSON
SCHOOL OF COMPUTING





Conclusions

CLEMSON
SCHOOL OF COMPUTING

- **Performance results show that KVM introduces a tolerable overhead for bag of tasks applications (<10%)**
- **Networking introduces significant latency that kills MPI applications.**
- **Need to investigate use of IPOP, ViNE or other Vnetwork.**

- **Early results for VOC dynamic behavior on the grid**
- **Work under way to build a VOC over WAN with own namespace.**

- **Ack: Work supported by NSF Grad fellowship, IBM and OSG**

Open Science Grid

CLEMSON
SCHOOL OF COMPUTING

Questions?

sebgoa@clemson.edu

http://cirg.cs.clemson.edu