Report on the Clusters at Fermilab

Don Holmgren

USQCD All-Hands Meeting

BNL

April 16, 2010

Outline

- Current Hardware
- Changes to Storage
- FY10/FY11 Deployment

Hardware – Current Clusters

Name	CPU	Nodes	Cores	Network	DWF	Asqtad	Online
Kaon	Dual 2.0 GHz Opteron 240 (Dual Core)	600	2400	Infiniband Double Data Rate	4696 MFlops per Node	3832 MFlops per Node	Oct 2006 2.56 TFlops
J/ψ	Dual 2.1 GHz Opteron 2352 (Quad Core)	856	6848	Infiniband Double Data Rate	10061 MFlops per Node	9563 MFlops per Node	Jan 2009 / Apr 2009 8.40 TFlops
Ds (2010)	Quad 2.0 GHz Opteron 6128 (8 Core) ???	240 ? ??	7680 ? ??	Infiniband Quad Data Rate	43 ?? GFlops per Node	48 ?? GFlops per Node	Nov 2010 11 TFlops

Hardware – GPUs

- Four Nvidia Tesla S1070 systems are available for CUDA programming and production
  - Each S1070 has 4 GPUs in 2 banks of 2
  - Each bank of 2 GPUs is attached to one dual Opteron node (32 GB of memory), accessed via the JPsi batch system
    - Nodes are “gpu01” through “gpu08”
    - Access via queue “gpu”
      (***`qsub -q gpu -l nodes=1 -I -A yourproject`***)
    - Parallel codes using multiple banks can use two or more nodes with MPI (or QMP) over Infiniband
    - Accounts are not charged for usage
  - Send mail to lqcd-admin@fnal.gov to request access

Hardware - Storage

- Current disk storage options:
  - 162 TB Lustre filesystem at /lqcdproj
  - 65 TB volatile dCache filesystem at /pnfs/volatile
  - 11.3 TB total NFS filesystems at /data/raid*x*
  - 3.1 TB total “project” space at /project (backed up nightly)
  - 6 GB per user at /home on each cluster (backed up nightly)
- Robotic tape storage is available via *dccp* commands against the dCache filesystem at /pnfs/lqcd

Storage – Planned Changes

1. Decommission /pnfs/volatile and redeploy as Lustre storage (+ 65 TB → 227 TB total) – July 1, 2010
   - Please give feedback:
     - will this affect your production?
     - how much data will you need to move from /pnfs/volatile to /lqcdproj?
2. Decommission /data/raid*x* – July 1, 2010
   - How much data will you need to move from /data/raid*x* to /lqcdproj?
3. Enforce group (project) quotas on /lqcdproj – July 1, 2010
   - Projects will be charged for disk and (new) tape usage at the beginning of each quarter
4. Deploy additional Lustre storage (+ ~144 TB → ~371 TB total) Sept 1, 2010

FY10/FY11 Deployment ("Ds")

- The LQCD-ext project has begun the combined FY10/FY11 purchase at Fermilab
- Configuration (most probable):
  - AMD-based ("Magny-Cours") dual- or quad-socket 8-core, or Intel-based ("Westmere") dual-socket quad-core
  - QDR Infiniband
- Expect friendly-user testing to begin September, and release in early November
- The FY10 portion (11 TF) will *not* have GPUs
- Some fraction of the FY11 funds will be used for GPUs (quantity and configuration TBD); the rest will be used to expand Ds

Performance of Current x86 Processors

Cluster	Processor	DWF Performance per Node	Clover Performance per Node	Asqtad Performance per Node
J/Psi	2.1 GHz Dual CPU Quad Core Opteron	10.1 GFlops	7.4 GFlops	9.6 GFlops
Intel Westmere	*2.53 GHz Dual CPU Quad Core Xeon*	*27.0 GFlops*	*13.4GFlops*	*16.6 GFlops*
AMD Magny-Cours	*2.0 GHz Dual CPU 8-Core Opteron*	*22.3 GFlops*	*17.4 GFlops*	*23.0 GFlops*
AMD Magny-Cours	*2.3 GHz Quad CPU 8-Core Opteron*	*45.1 GFlops*	*35.1 GFlops*	*49.6 GFlops*

- J/Psi performance figures are from 128-process parallel runs (90% scaling from single to 16-nodes)

- *Westmere and Magny-Cours performance figures are estimated from measured single node performance using a conservative 85% scaling factor (we are more likely to see 90%)*

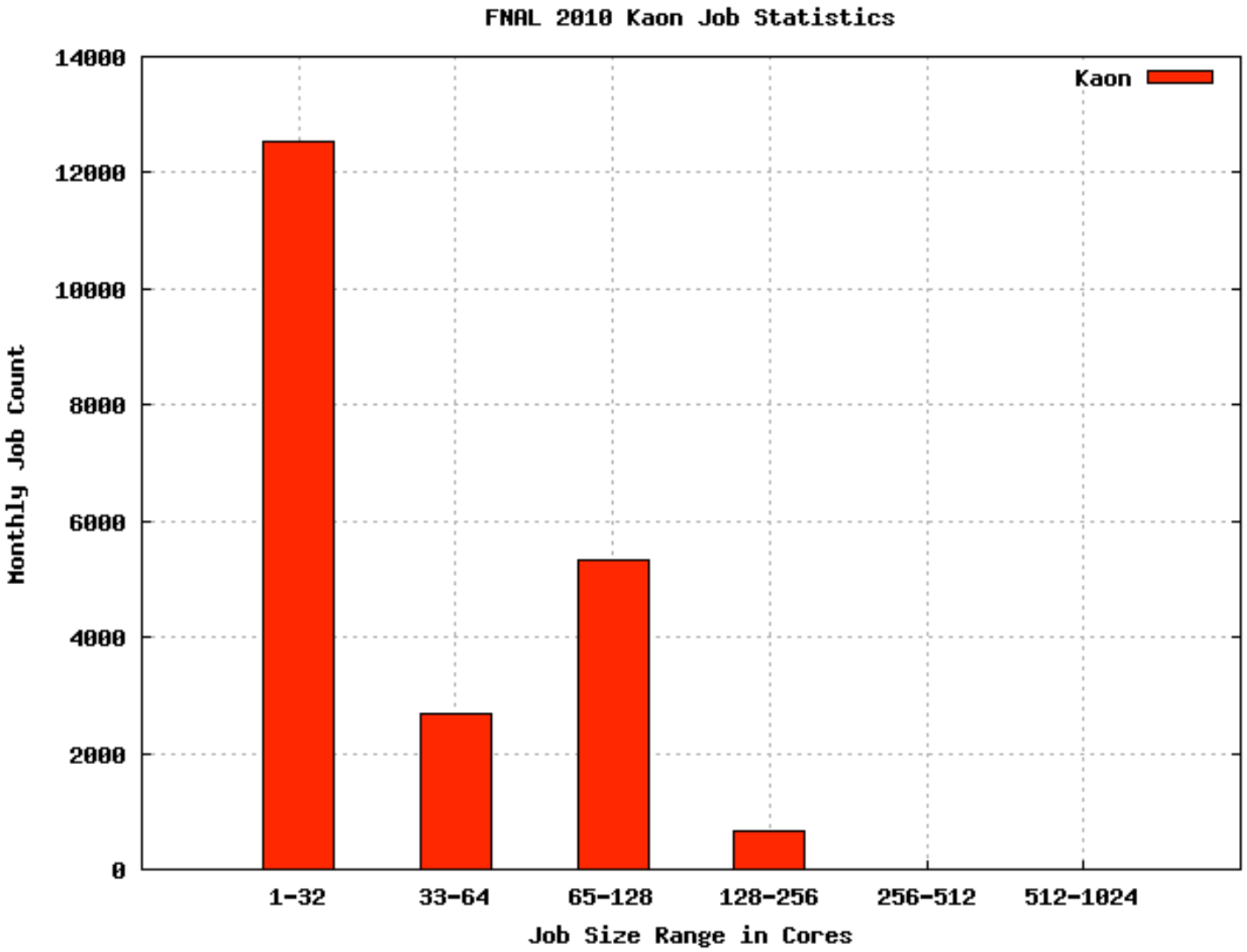
Ds Questions

- There is a strong possibility that Ds will be based on 32-core nodes
  - Is 64 GB of memory per node (2 GB/core) sufficient?
  - Are there production streams that *cannot* take advantage of 32 cores?

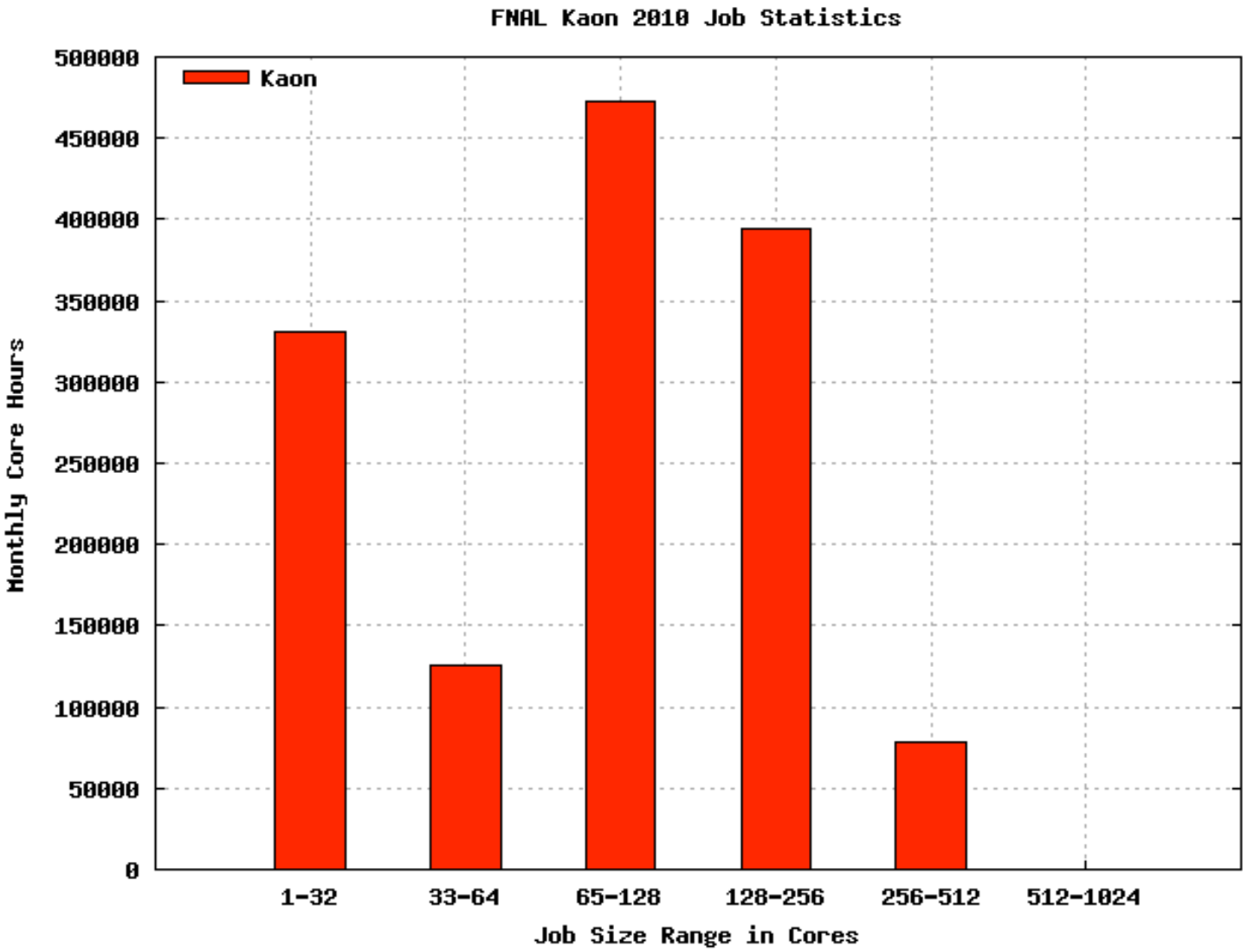
Statistics

- Since April 1, 2008, including QCD, Pion, Kaon, JPsi:
  - Users submitting jobs:
    63 USQCD, 9 administrators or other
  - 1,173,767 jobs (857,474 multi-node)
  - 14.0M node-hours → 34.7M 6n-node-hours = 68.8M JPsi-core-hours
    (7200-hr year capacity: 35.5M 6n-node-hours = 70.3M JPsi-core-hrs)

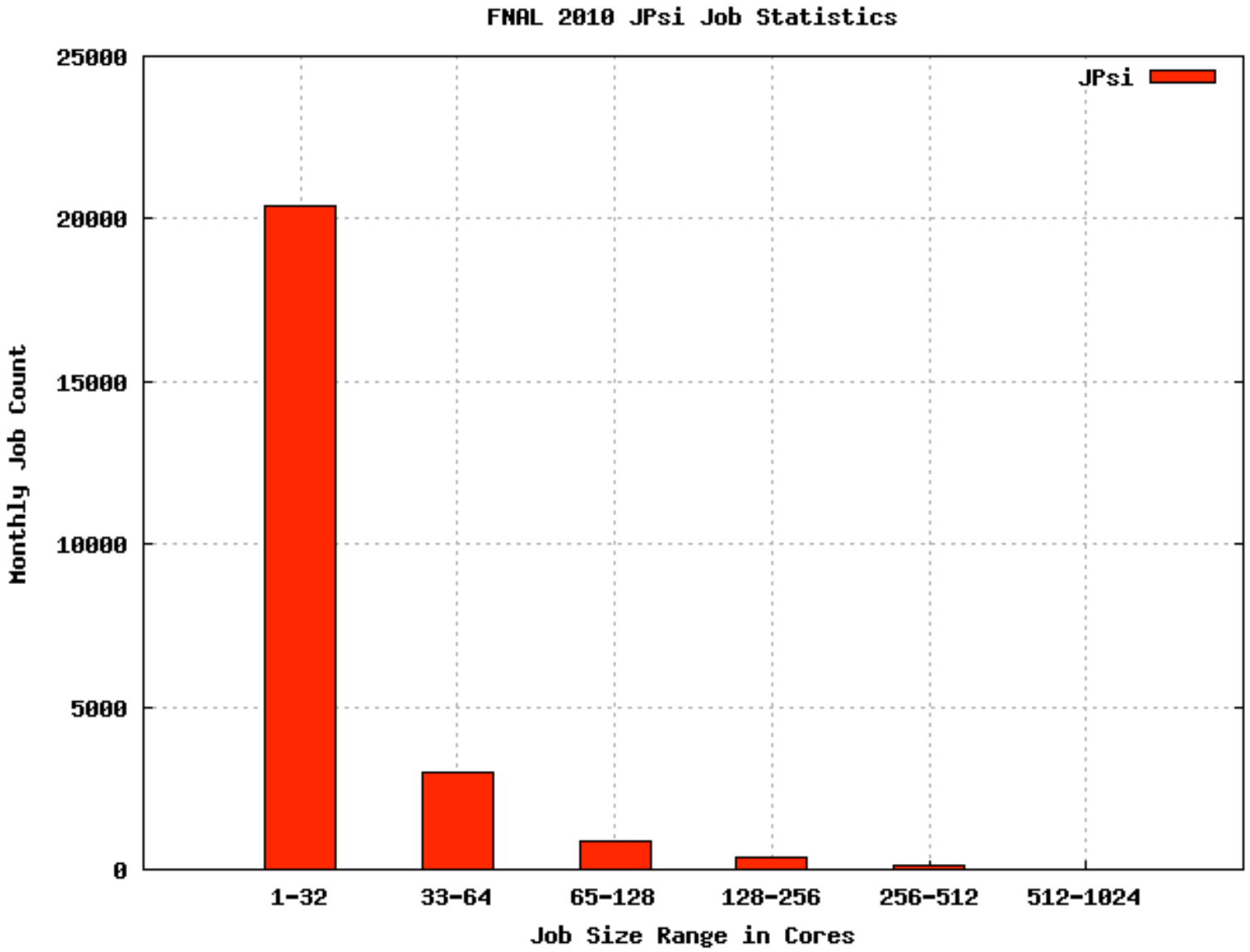
Kaon Job Statistics



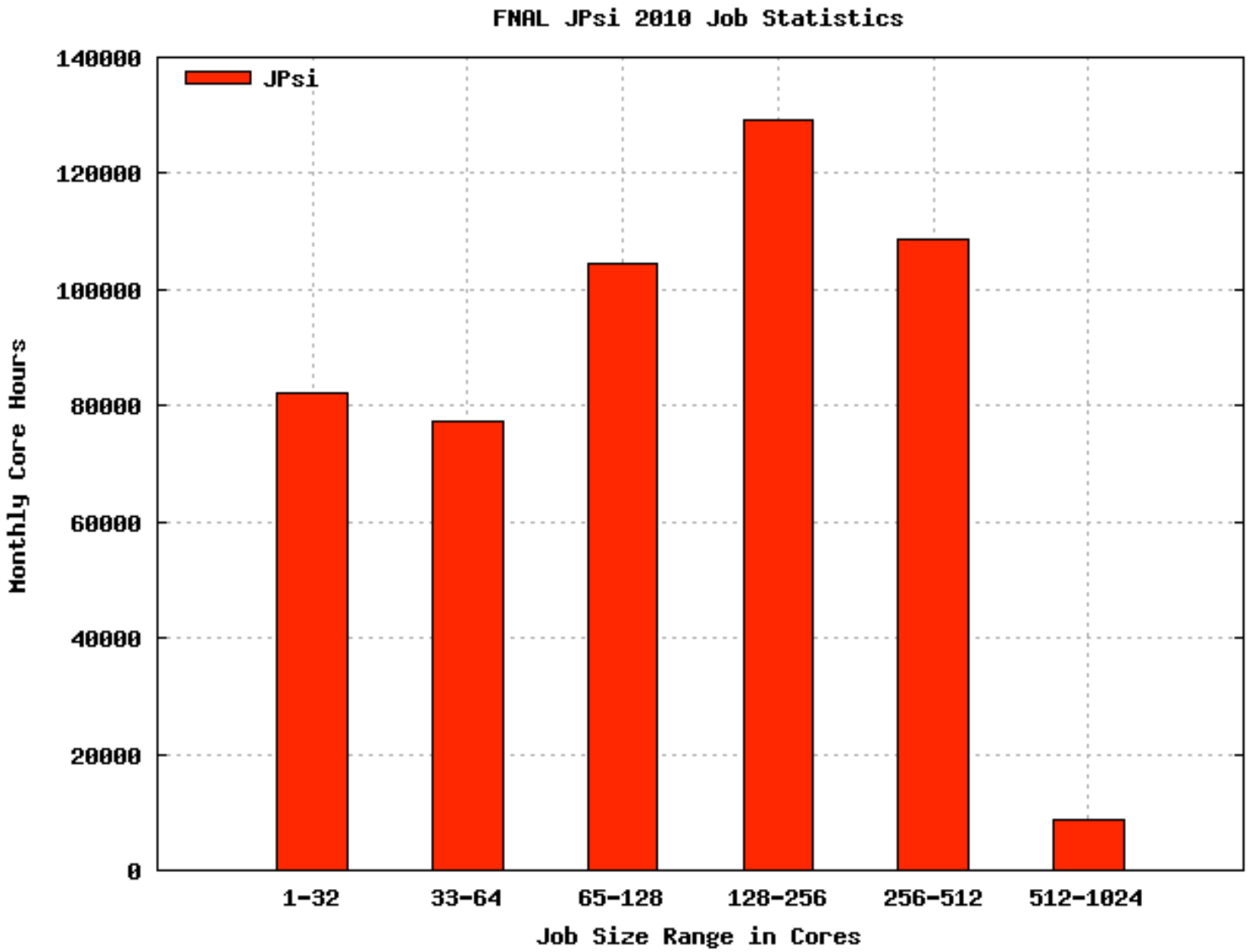
Kaon Core Hour Statistics



JPsi Job Statistics



JPsi Core Hour Statistics



User Support

Fermilab points of contact:

- Best choice: lqcd-admin@fnal.gov
- Don Holmgren, djholm@fnal.gov
- Amitoj Singh, amitoj@fnal.gov
- Kurt Ruthmansdorfer, kurt@fnal.gov
- Nirmal Seenu, nirmal@fnal.gov
- Jim Simone, simone@fnal.gov
- Ken Schumacher, kschu@fnal.gov
- Rick van Conant, vanconant@fnal.gov
- Bob Forster, forster@fnal.gov
- Paul Mackenzie, mackenzie@fnal.gov