



THE UNIVERSITY OF  
**CHICAGO**



## Running 1 Million Jobs in 10 Minutes via the Falkon Fast and Light-weight task executiON framework

**Ioan Raicu**

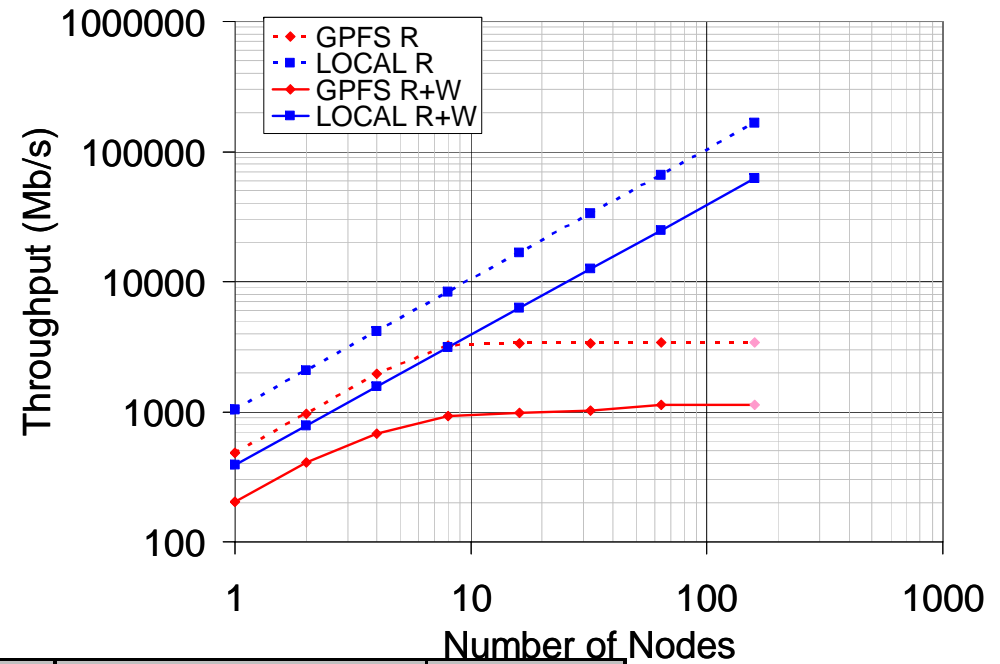
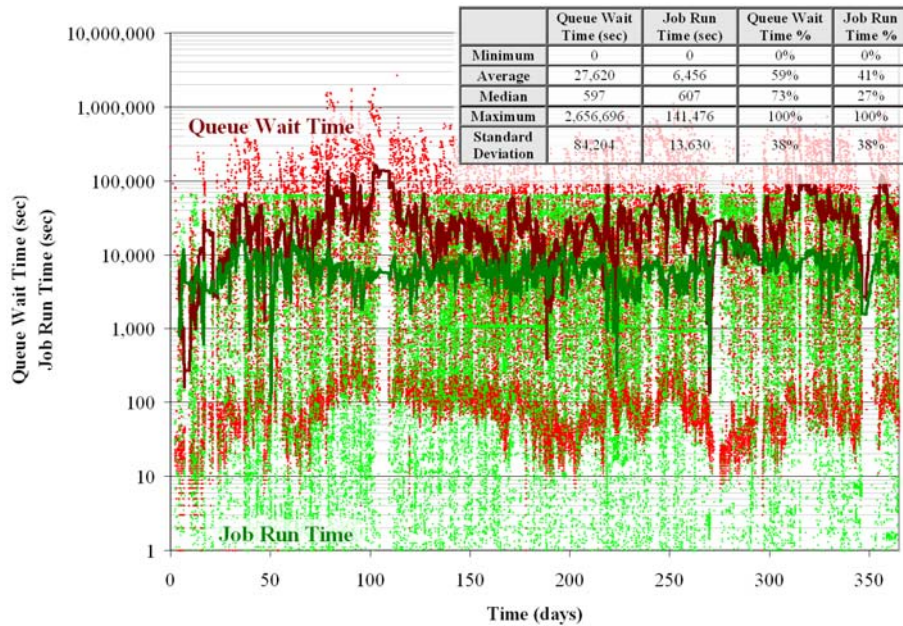
Distributed Systems Laboratory  
Computer Science Department  
University of Chicago

**In Collaboration with:**

**Ian Foster**, University of Chicago and Argonne National Laboratory  
**Mike Wilde**, University of Chicago and Argonne National Laboratory  
**Zhao Zhang**, University of Chicago  
**Yong Zhao**, Microsoft  
**Catalin Dumitrescu**, Fermi National Laboratory

Megajob BOF at IEEE/ACM Supercomputing 2008  
November 18<sup>th</sup>, 2008

# Obstacles running MTC apps in Clusters/Grids



System	Comments	Throughput (tasks/sec)
<b>Condor (v6.7.2) - Production</b>	Dual Xeon 2.4GHz, 4GB	0.49
<b>PBS (v2.1.8) - Production</b>	Dual Xeon 2.4GHz, 4GB	0.45
<b>Condor (v6.7.2) - Production</b>	Quad Xeon 3 GHz, 4GB	2
<b>Condor (v6.8.2) - Production</b>		0.42
<b>Condor (v6.9.3) - Development</b>		11
<b>Condor-J2 - Experimental</b>	Quad Xeon 3 GHz, 4GB	22

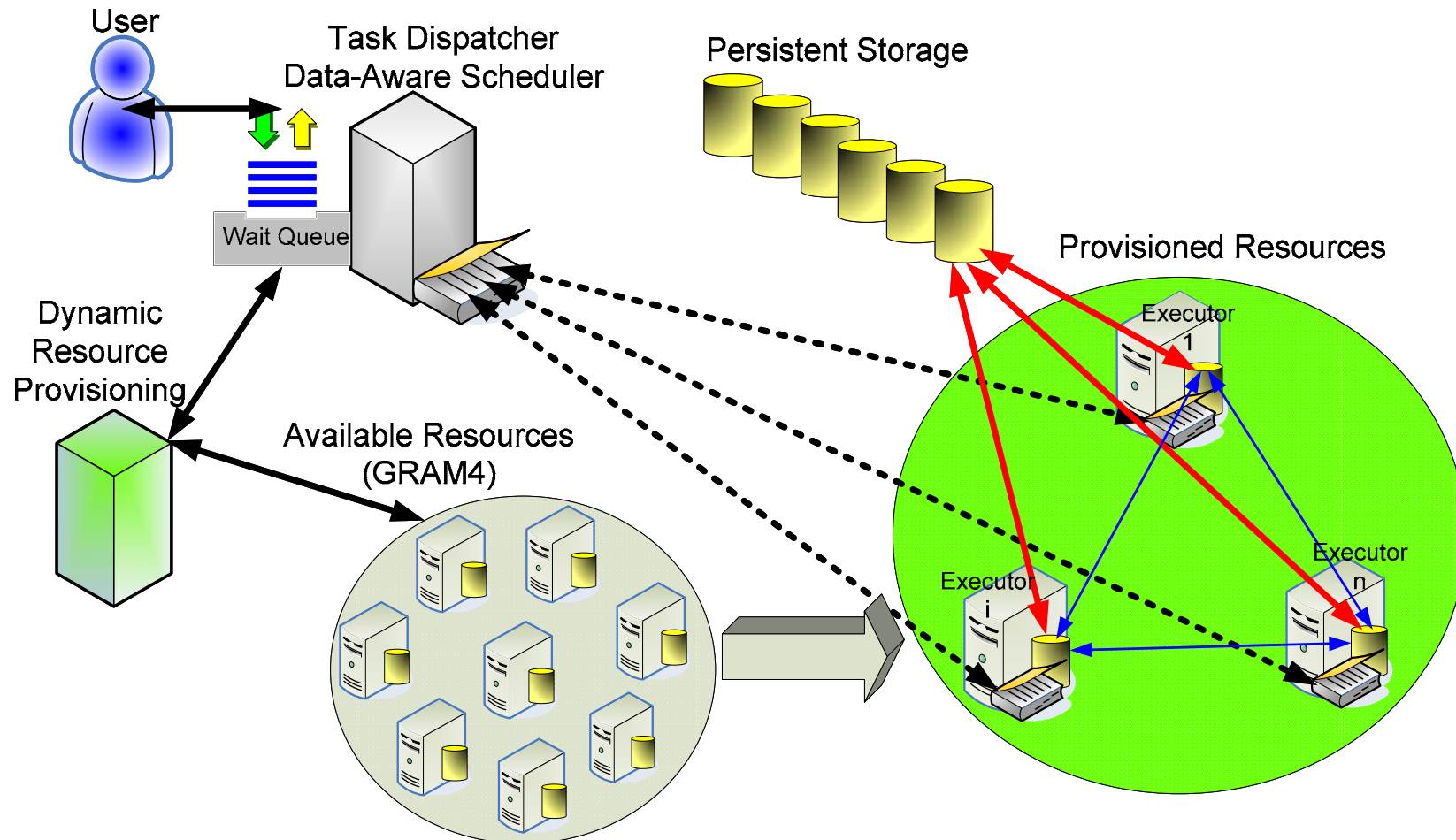
Running 1 Million Jobs in 10 Minutes via the Falcon Fast and Light-weight task execution framework

# Solution



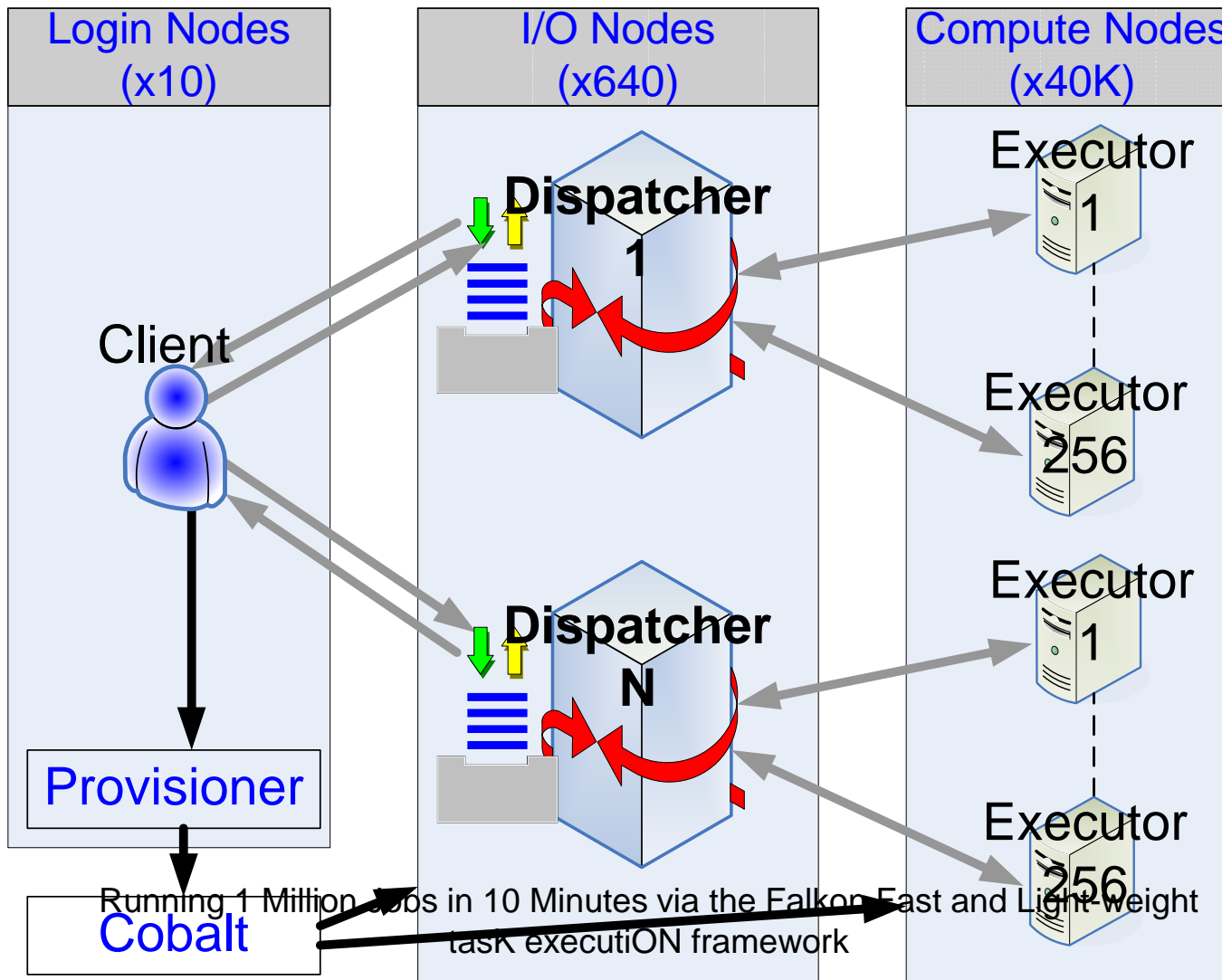
- Falkon: A Fast and Light-weight task executiON framework
  - **Goal:** enable the *rapid and efficient* execution of many independent jobs on large compute clusters
  - Combines three components:
    - A *streamlined task dispatcher*
    - **Resource provisioning** through multi-level scheduling techniques
    - **Data diffusion** and data-aware scheduling to leverage the co-located computational and storage resources

# Falkon Overview

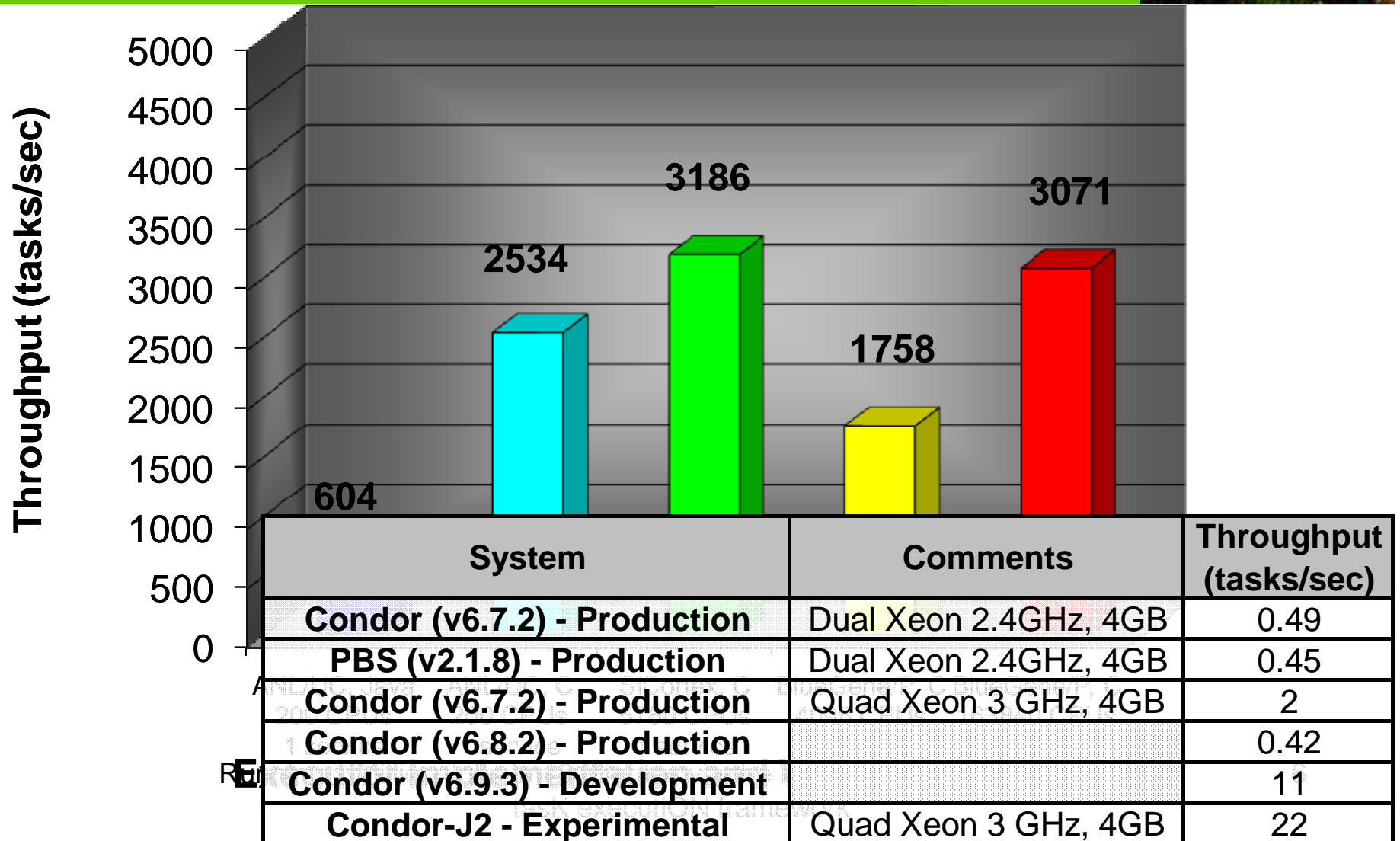


Running 1 Million Jobs in 10 Minutes via the Falkon Fast and Light-weight task execution framework

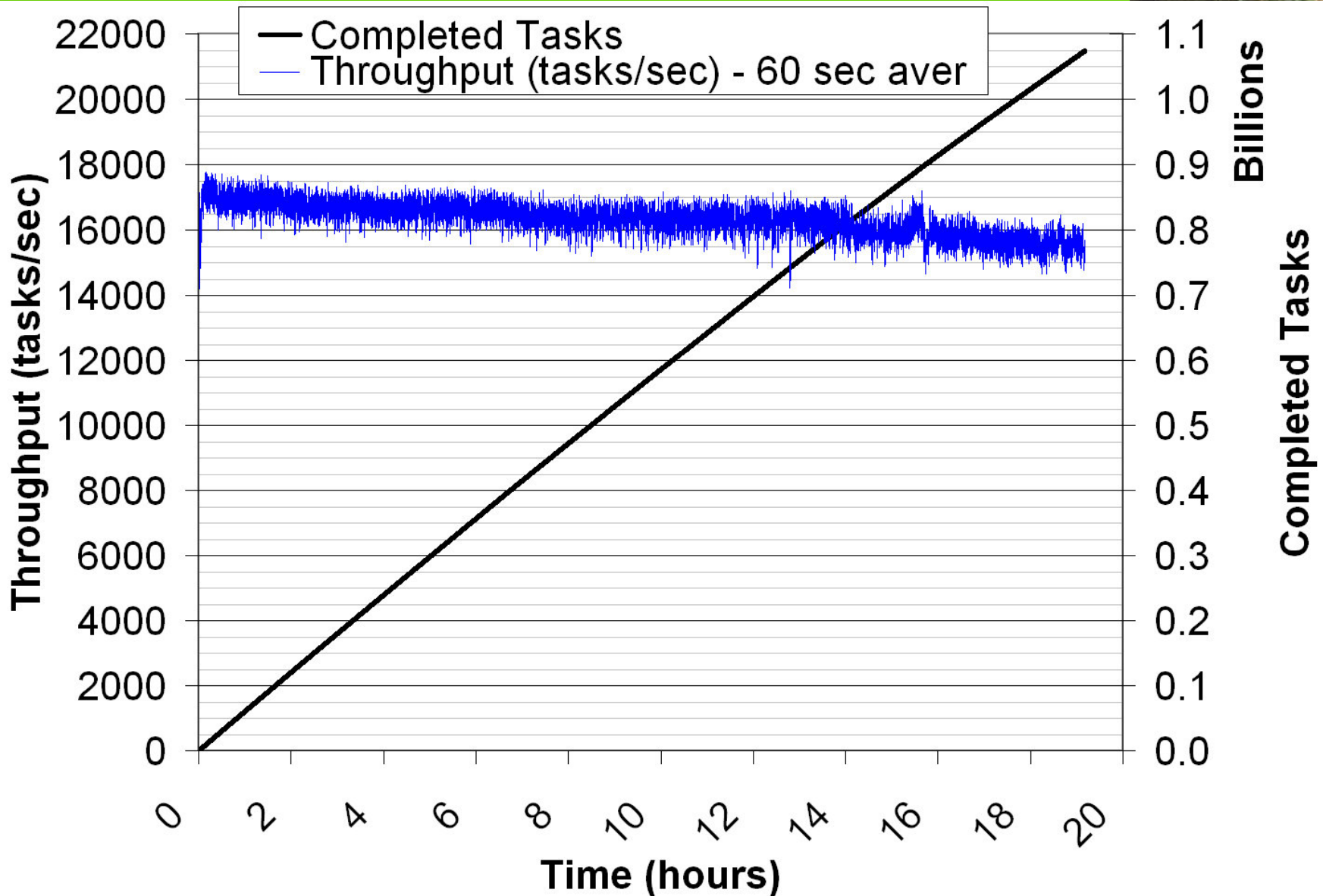
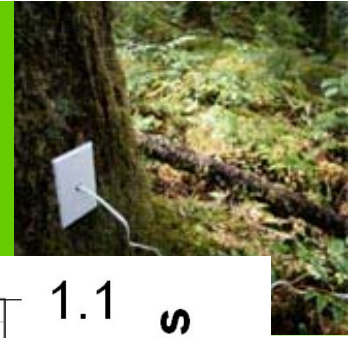
# Distributed Falcon Architecture



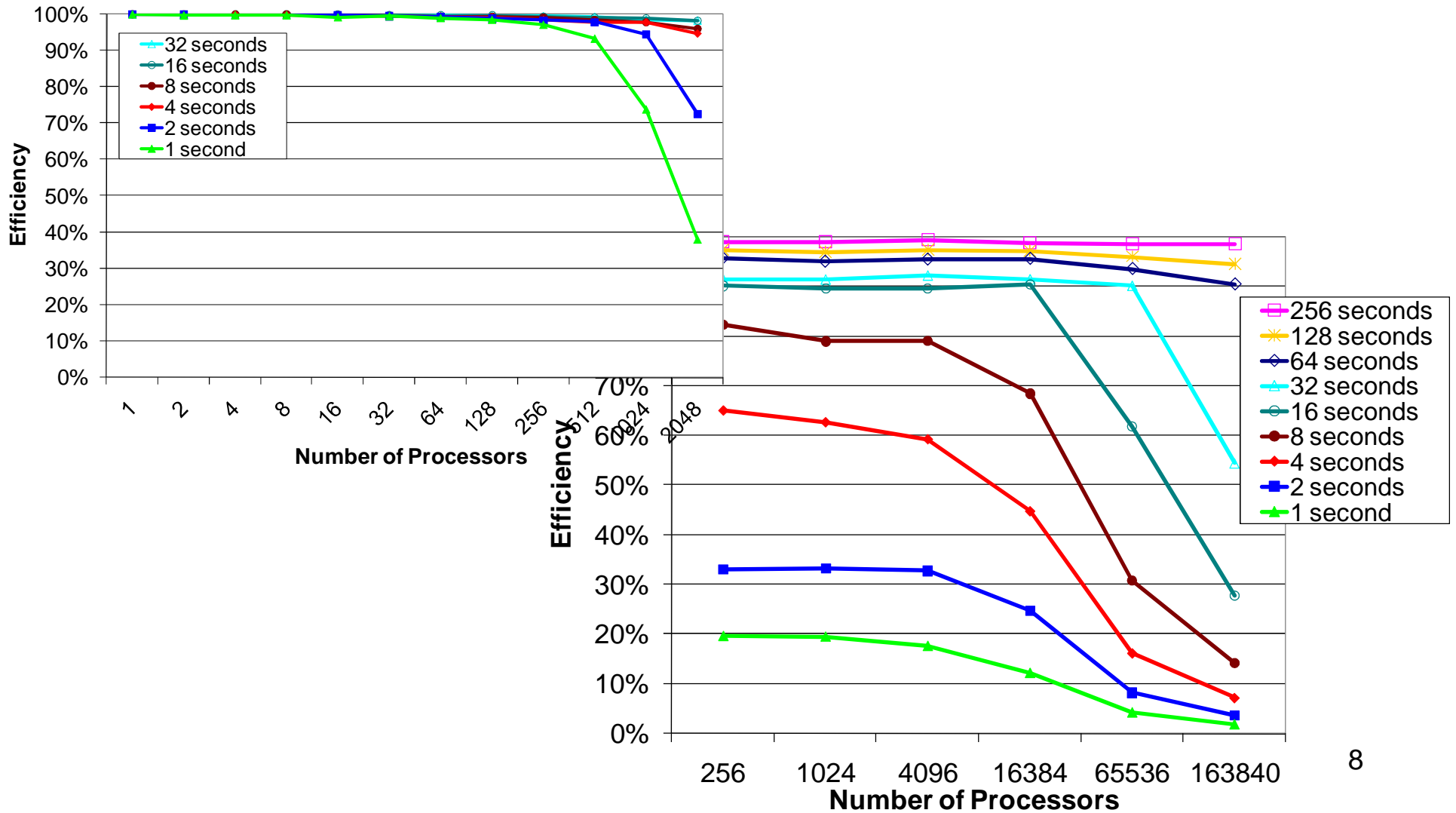
# Dispatch Throughput



# Falkon Endurance Test

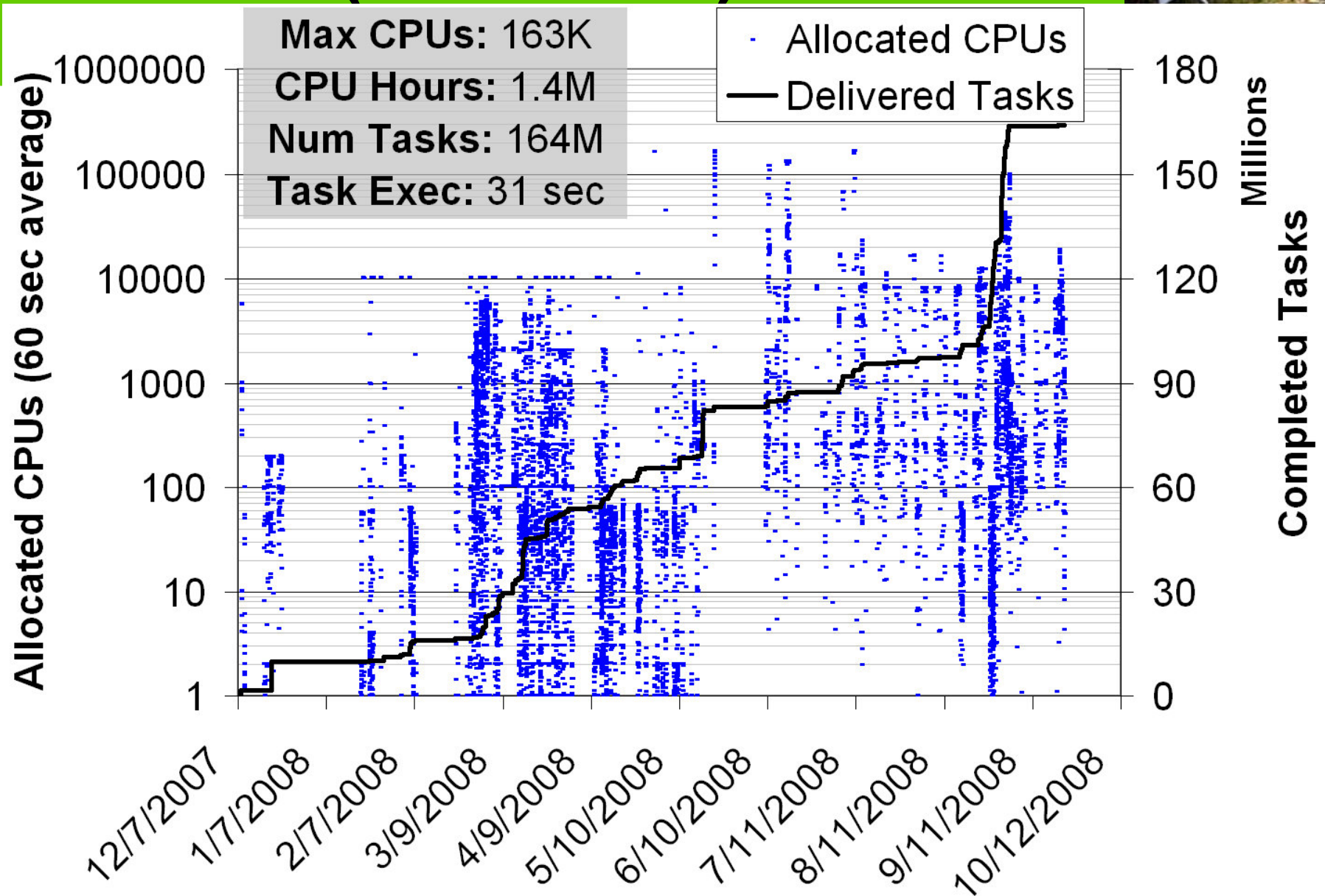
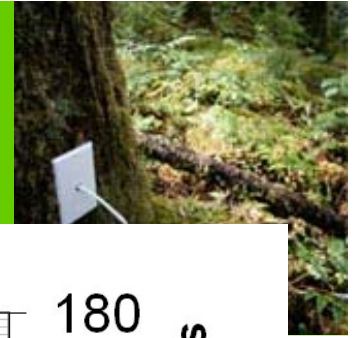


# Efficiency





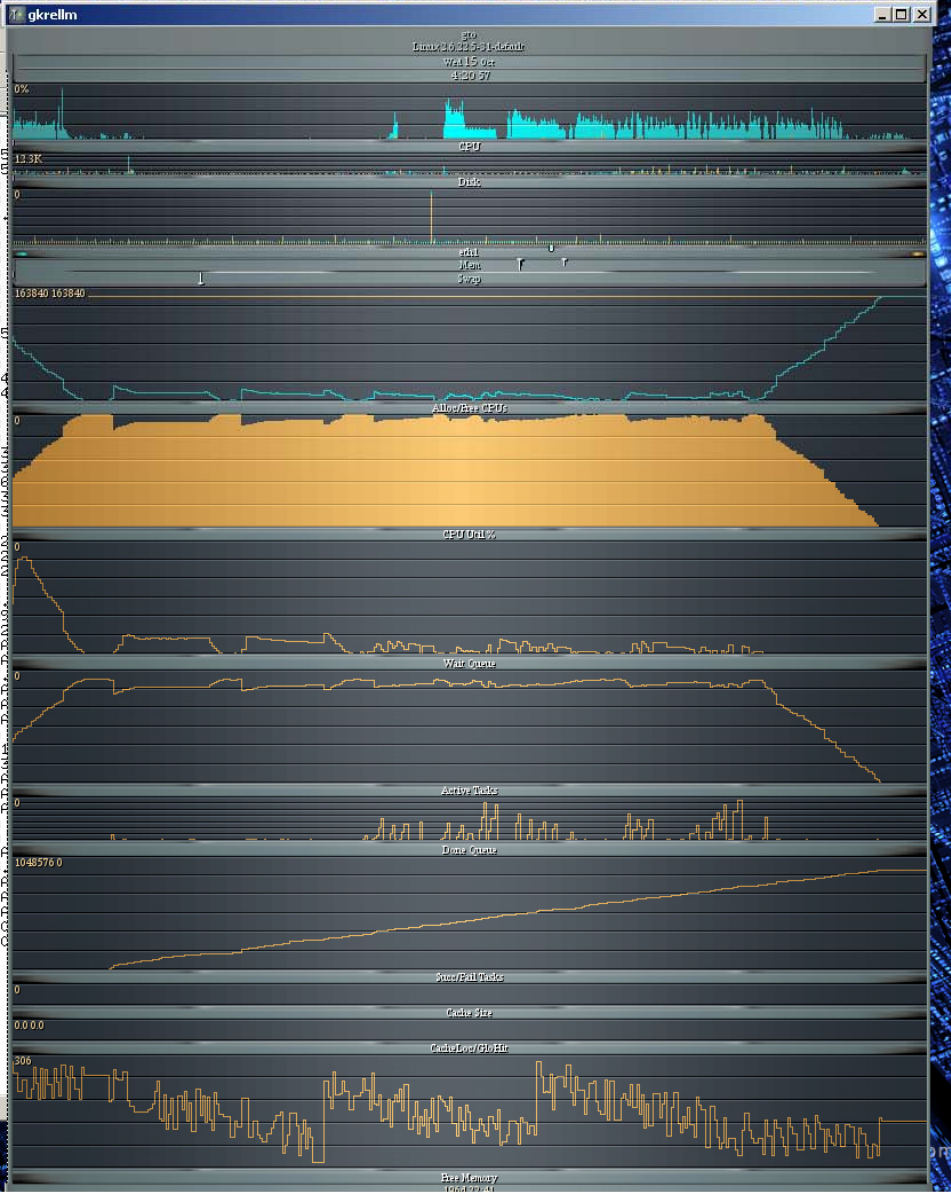
# Falkon Activity History (10 months)



# Falkon Demo

```
gto.ci.uchicago.edu (1) - SecureCRT
File Edit View Options Transfer Script Tools Help
gto.ci.uchicago.edu | gto.ci.uchicago.edu (1) | gto.ci.uchicago.edu (3) | gto.ci.uchicago.edu (2) | gto.ci.uchicago.edu (5) | gto.ci.uchicago.edu (4)
397,951 tasks+ 908675 tasks- 0 tasks-> 1048576 completed 86.66 tasks_tp 3246.03 aver_tp 2695.68 stdev_tp 3157.365 ETA
398,959 tasks+ 911918 tasks- 0 tasks-> 1048576 completed 86.97 tasks_tp 3217.26 aver_tp 2697.24 stdev_tp 3152.763 ETA
399,967 tasks+ 913940 tasks- 0 tasks-> 1048576 completed 87.16 tasks_tp 3205.95 aver_tp 2695.18 stdev_tp 3148.28 ETA
400,975 tasks+ 916630 tasks- 0 tasks-> 1048576 completed 87.42 tasks_tp 3268.65 aver_tp 2695.1 stdev_tp 3143.592 ETA
401,984 tasks+ 919282 tasks- 0 tasks-> 1048576 completed 87.67 tasks_tp 3230.95 aver_tp 2694.91 stdev_tp 3138.926 ETA
402,992 tasks+ 921616 tasks- 0 tasks-> 1048576 completed 87.89 tasks_tp 3215.48 aver_tp 2693.79 stdev_tp 3134.347 ETA
404,0 tasks+ 924266 tasks- 0 tasks-> 1048576 completed 88.14 tasks_tp 2628.97 aver_tp 2693.6 stdev_tp 3129.723 ETA
405,004 tasks+ 926864 tasks- 0 tasks-> 1048576 completed 88.39 tasks_tp 2587.65 aver_tp 2693.29 stdev_tp 3125.122 ETA
406,008 tasks+ 929627 tasks- 0 tasks-> 1048576 completed 88.66 tasks_tp 2751.99 aver_tp 2693.46 stdev_tp 3120.538 ETA
407,013 tasks+ 932059 tasks- 0 tasks-> 1048576 completed 89.09 tasks_tp 2422.31 aver_tp 2692.65 stdev_tp 3116.007 ETA
408,017 tasks+ 934610 tasks- 0 tasks-> 1048576 completed 89.13 tasks_tp 2540.84 aver_tp 2692.22 stdev_tp 3111.472 ETA
409,021 tasks+ 937289 tasks- 0 tasks-> 1048576 completed 89.36 tasks_tp 2439.24 aver_tp 2691.49 stdev_tp 3106.976 ETA
410,025 tasks+ 939999 tasks- 0 tasks-> 1048576 completed 89.57 tasks_tp 2122.51 aver_tp 2689.84 stdev_tp 3102.621 ETA
411,029 tasks+ 942746 tasks- 0 tasks-> 1048576 completed 89.75 tasks_tp 2279.88 aver_tp 2690.65 stdev_tp 3098.212 ETA
412,033 tasks+ 945526 tasks- 0 tasks-> 1048576 completed 90.0 tasks_tp 2218.13 aver_tp 2687.3 stdev_tp 3093.948 ETA
413,038 tasks+ 948346 tasks- 0 tasks-> 1048576 completed 90.21 tasks_tp 2171.31 aver_tp 2695.81 stdev_tp 3089.523 ETA
414,042 tasks+ 949125 tasks- 0 tasks-> 1048576 completed 90.42 tasks_tp 2234.06 aver_tp 2694.52 stdev_tp 3085.188 ETA
415,046 tasks+ 950185 tasks- 0 tasks-> 1048576 completed 90.62 tasks_tp 2047.81 aver_tp 2692.7 stdev_tp 3080.965 ETA
416,051 tasks+ 952338 tasks- 0 tasks-> 1048576 completed 90.81 tasks_tp 2144.42 aver_tp 2681.17 stdev_tp 3076.707 ETA
417,054 tasks+ 954561 tasks- 0 tasks-> 1048576 completed 91.0 tasks_tp 2214.14 aver_tp 2679.84 stdev_tp 3072.434 ETA
418,062 tasks+ 955645 tasks- 0 tasks-> 1048576 completed 91.23 tasks_tp 2067.46 aver_tp 2678.11 stdev_tp 3068.251 ETA
419,071 tasks+ 956742 tasks- 0 tasks-> 1048576 completed 91.43 tasks_tp 2090.36 aver_tp 2676.42 stdev_tp 3064.079 ETA
420,079 tasks+ 957890 tasks- 0 tasks-> 1048576 completed 91.6 tasks_tp 1724.21 aver_tp 2673.73 stdev_tp 3060.176 ETA
421,087 tasks+ 958605 tasks- 0 tasks-> 1048576 completed 91.8 tasks_tp 2108.13 aver_tp 2672.15 stdev_tp 3056.022 ETA
422,095 tasks+ 959457 tasks- 0 tasks-> 1048576 completed 92.0 tasks_tp 2075.4 aver_tp 2670.47 stdev_tp 3051.902 ETA
423,103 tasks+ 959960 tasks- 0 tasks-> 1048576 completed 92.12 tasks_tp 1248.02 aver_tp 2666.5 stdev_tp 3048.961 ETA
424,111 tasks+ 974461 tasks- 0 tasks-> 1048576 completed 92.93 tasks_tp 8425.6 aver_tp 2682.54 stdev_tp 3059.406 ETA
425,119 tasks+ 978213 tasks- 0 tasks-> 1048576 completed 93.23 tasks_tp 3722.22 aver_tp 2685.43 stdev_tp 3065.644 ETA
426,128 tasks+ 980343 tasks- 0 tasks-> 1048576 completed 93.48 tasks_tp 3011.9 aver_tp 2683.57 stdev_tp 3061.614 ETA
427,136 tasks+ 982449 tasks- 0 tasks-> 1048576 completed 93.77 tasks_tp 437.5 aver_tp 2682.19 stdev_tp 3047.598 ETA
428,144 tasks+ 983411 tasks- 0 tasks-> 1048576 completed 94.02 tasks_tp 347.31 aver_tp 2677.45 stdev_tp 3044.643 ETA
429,152 tasks+ 987523 tasks- 0 tasks-> 1048576 completed 94.22 tasks_tp 2044.76 aver_tp 2691.51 stdev_tp 3041.641 ETA
430,161 tasks+ 989220 tasks- 0 tasks-> 1048576 completed 94.42 tasks_tp 232.26 aver_tp 2694.57 stdev_tp 3048.1 ETH 23
431,169 tasks+ 993260 tasks- 0 tasks-> 1048576 completed 94.92 tasks_tp 0.0 aver_tp 2687.6 stdev_tp 3047.182 ETA
432,176 tasks+ 997217 tasks- 0 tasks-> 1048576 completed 95.1 tasks_tp 1941.47 aver_tp 2685.57 stdev_tp 3043.276 ETA
433,184 tasks+ 999111 tasks- 0 tasks-> 1048576 completed 95.57 tasks_tp 1879.97 aver_tp 2691.53 stdev_tp 3041.282 ETA
434,193 tasks+ 1000776 tasks- 0 tasks-> 1048576 completed 95.99 tasks_tp 191.53 aver_tp 2686.7 stdev_tp 3040.335 ETA
435,201 tasks+ 1000776 tasks- 0 tasks-> 1048576 completed 96.47 tasks_tp 263.26 aver_tp 2691.26 stdev_tp 3044.495 ETA
436,209 tasks+ 1004983 tasks- 0 tasks-> 1048576 completed 96.93 tasks_tp 1527.58 aver_tp 2688.38 stdev_tp 3031.597 ETA
437,217 tasks+ 1011812 tasks- 0 tasks-> 1048576 completed 96.49 tasks_tp 1527.58 aver_tp 2688.38 stdev_tp 3031.597 ETA
438,225 tasks+ 1015395 tasks- 0 tasks-> 1048576 completed 96.84 tasks_tp 3555.56 aver_tp 2690.71 stdev_tp 3027.863 ETA
439,233 tasks+ 1019585 tasks- 0 tasks-> 1048576 completed 97.27 tasks_tp 4850.6 aver_tp 2695.04 stdev_tp 3025.337 ETA
440,241 tasks+ 1024499 tasks- 0 tasks-> 1048576 completed 97.77 tasks_tp 4850.6 aver_tp 2695.04 stdev_tp 3025.337 ETA
441,249 tasks+ 1029031 tasks- 0 tasks-> 1048576 completed 98.25 tasks_tp 4472.22 aver_tp 2694.02 stdev_tp 3016.048 ETA
442,257 tasks+ 1032856 tasks- 0 tasks-> 1048576 completed 98.47 tasks_tp 2313.64 aver_tp 2693.04 stdev_tp 3012.127 ETA
443,266 tasks+ 1036223 tasks- 0 tasks-> 1048576 completed 98.83 tasks_tp 3662.7 aver_tp 2695.59 stdev_tp 3009.572 ETA
444,274 tasks+ 1039578 tasks- 0 tasks-> 1048576 completed 99.17 tasks_tp 2584.48 aver_tp 2695.24 stdev_tp 3004.628 ETA
445,282 tasks+ 1036258 tasks- 0 tasks-> 1048576 completed 99.25 tasks_tp 1923.56 aver_tp 2693.24 stdev_tp 3000.949 ETA
446,291 tasks+ 1041293 tasks- 0 tasks-> 1048576 completed 99.3 tasks_tp 395.83 aver_tp 2687.24 stdev_tp 2999.32 ETA
447,300 tasks+ 1044203 tasks- 0 tasks-> 1048576 completed 99.51 tasks_tp 2265.87 aver_tp 2686.15 stdev_tp 2995.499 ETA
448,308 tasks+ 1046203 tasks- 0 tasks-> 1048576 completed 99.77 tasks_tp 2683.18 aver_tp 2686.15 stdev_tp 2991.596 ETA
449,314 tasks+ 1042883 tasks- 0 tasks-> 1048576 completed 100.0 tasks_tp 2354.17 aver_tp 2685.29 stdev_tp 2987.766 ETA
450,322 tasks+ 1046203 tasks- 0 tasks-> 1048576 completed 100.0 tasks_tp 0.0 aver_tp 2678.35 stdev_tp 2987.016 ETA
451,331 tasks+ 1048576 tasks- 0 tasks-> 1048576 completed 100.0 tasks_tp 0.0 aver_tp 2671.45 stdev_tp 2986.253 ETA
452,339 tasks+ 1048576 tasks- 0 tasks-> 1048576 completed 100.0 tasks_tp 0.0 aver_tp 2671.45 stdev_tp 2986.253 ETA
453,347 tasks+ 1048576 tasks- 0 tasks-> 1048576 completed 100.0 tasks_tp 0.0 aver_tp 2671.45 stdev_tp 2986.253 ETA
1048576 tasks completed in 453.505 sec
Successful tasks: 1048576
Failed tasks: 0
Notification Errors: 0
Overall Throughput (tasks/sec): 2312.16
Overall Throughput Standard Deviation: 2986.253
waiting to destroy all resources...
ShutdownHook triggered successfully!
iraicu@gto:~/falkon>
```

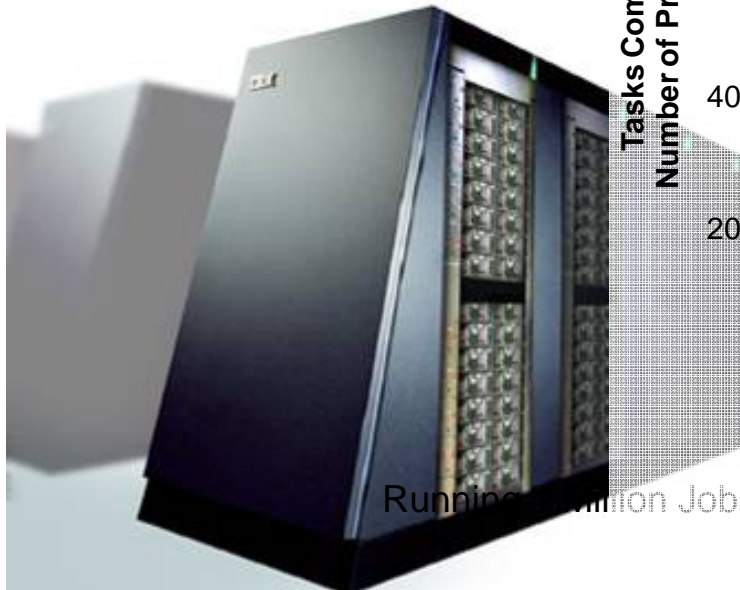
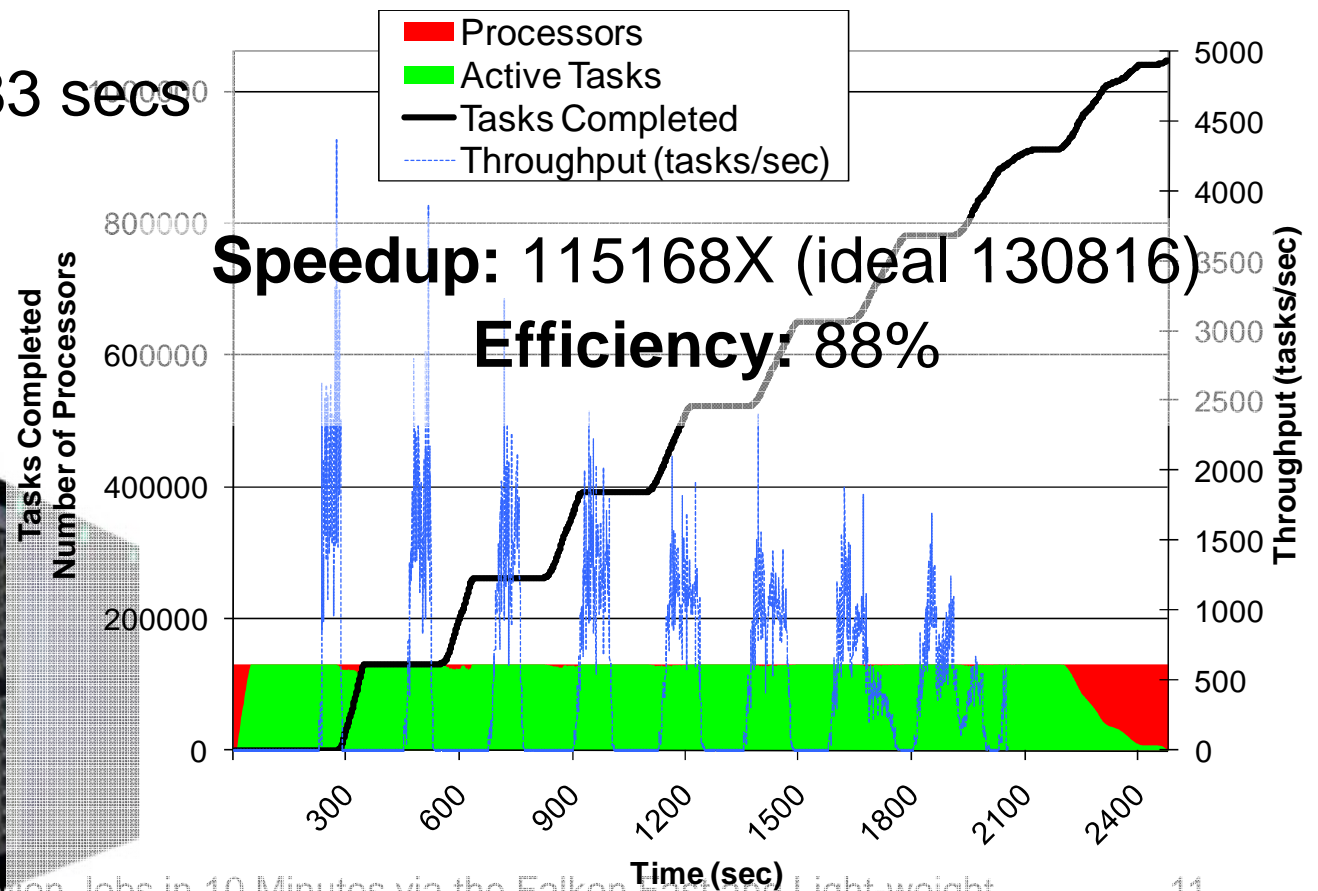
- Workload
- 160K CPUs
- 1M tasks
- 60 sec per task
- 17.5K CPU hours in 7.5 min
- Throughput: 2312 tasks/sec
- 85% efficiency



# MARS Economic Modeling on IBM BG/P (128K CPUs)



- CPU Cores: 130816
- Tasks: 1048576
- Elapsed time: 2483 secs
- CPU Years: 9.3



Running 10 Million Jobs in 10 Minutes via the Falcon Fast and Light-weight task execution framework

# DOCK on the BG/P



CPU cores: 118784

Tasks: 934803

Elapsed time: 2.01 hours

Compute time: 21.43 CPU years

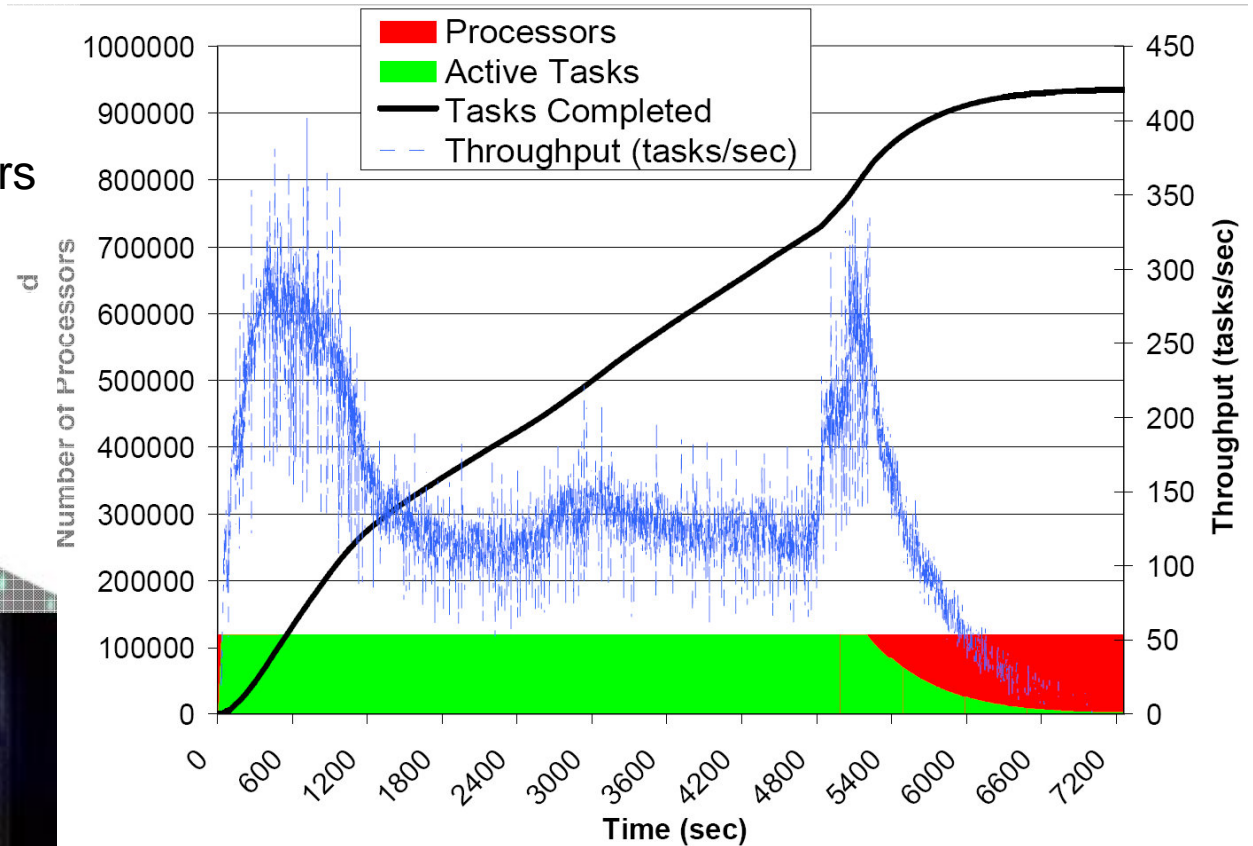
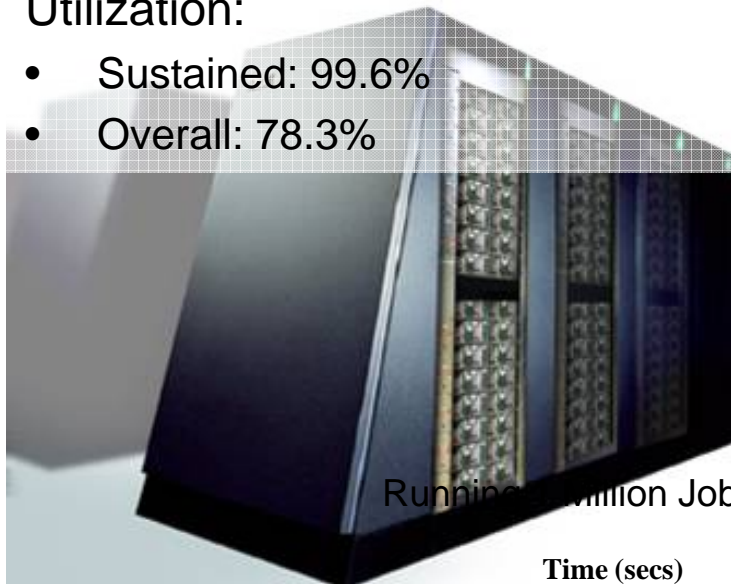
Average task time: 667 sec

Relative Efficiency: 99.7%

(from 16 to 32 racks)

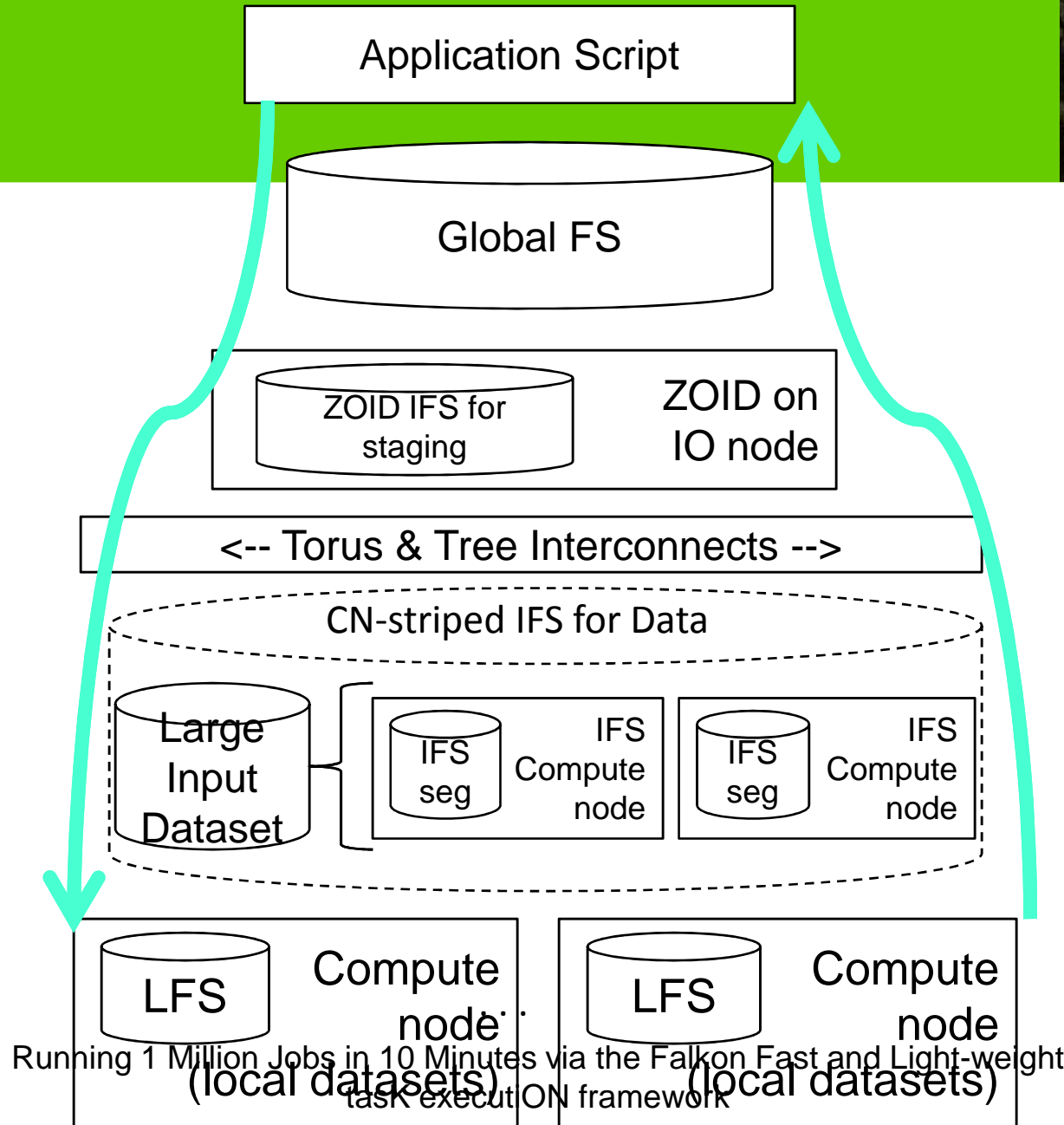
Utilization:

- Sustained: 99.6%
- Overall: 78.3%



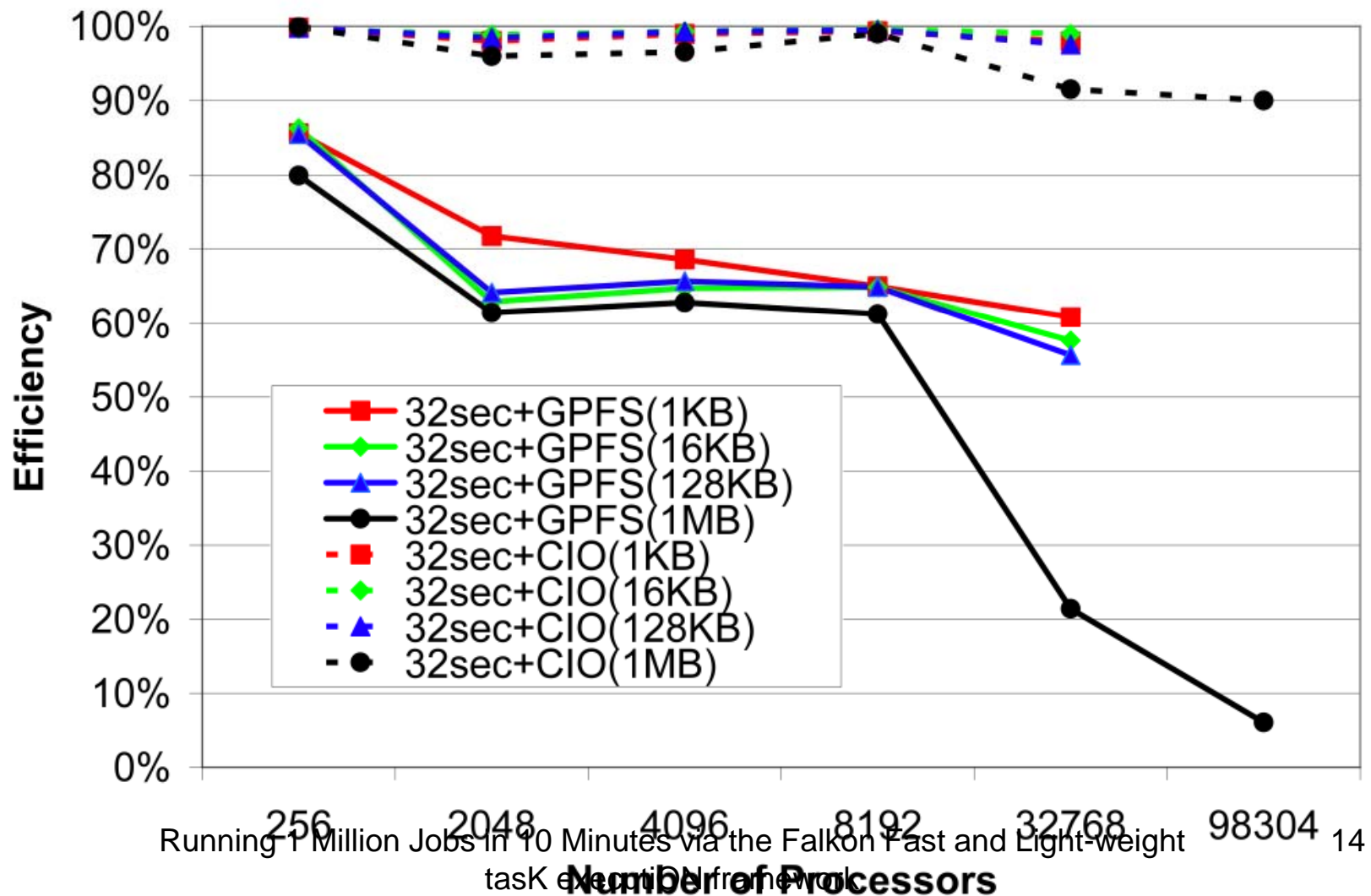
Running 1 Million Jobs in 10 Minutes via the Falcon Fast and Light-weight task execution framework

# Collective IO Model



# Write Performance

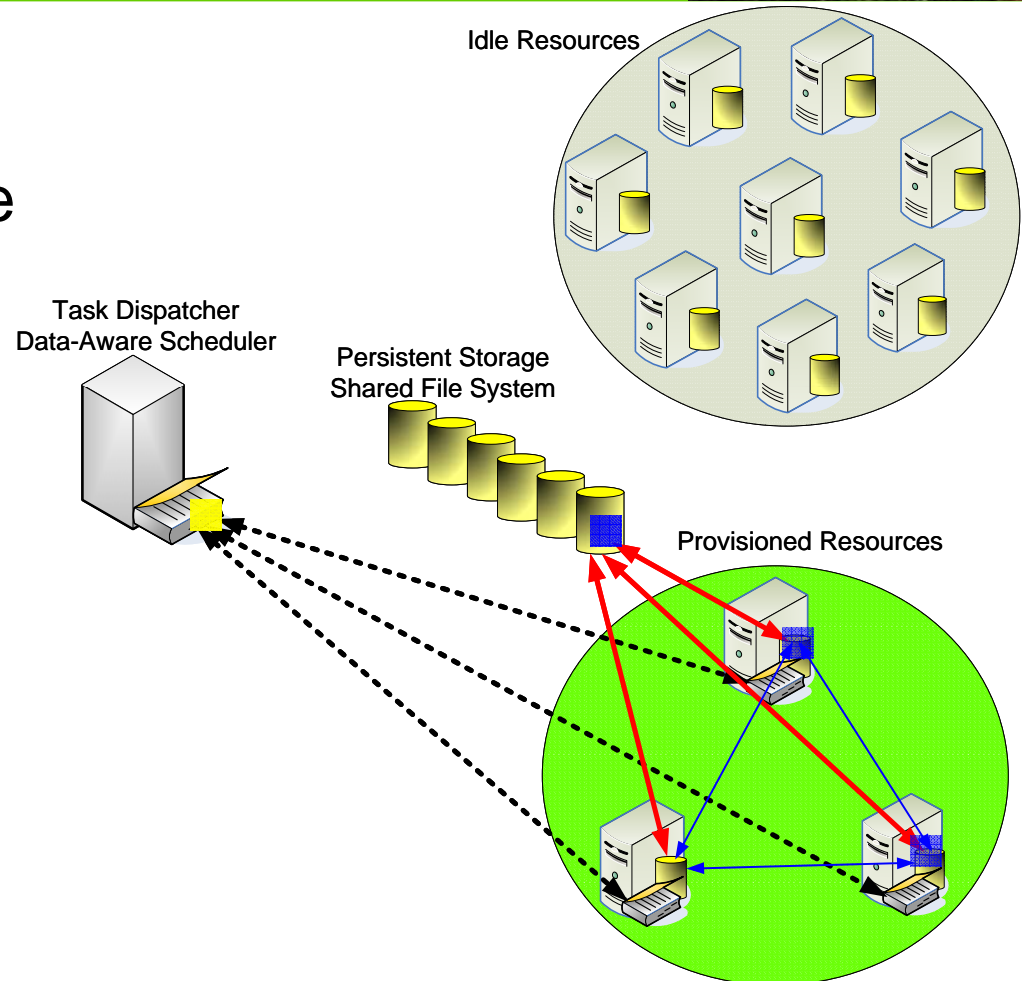
## CIO vs. GPFS efficiency



# Data Diffusion

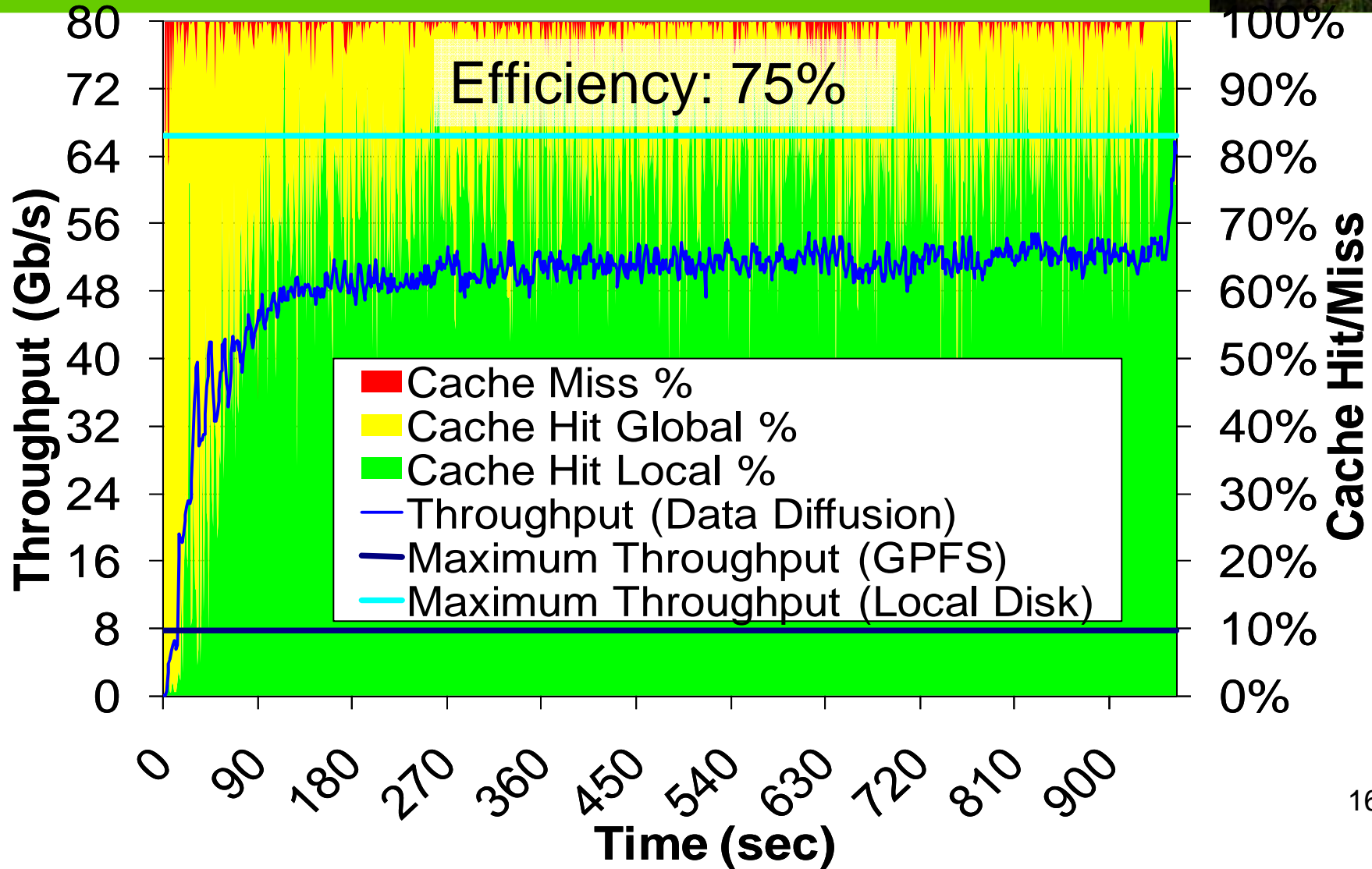


- Resource acquired in response to demand
- Data and applications diffuse from archival storage to newly acquired resources
- Resource “caching” allows faster responses to subsequent requests
  - Cache Eviction Strategies: RANDOM, FIFO, LRU, LFU
- Resources are released when demand drops



Running 1 Million Jobs in 10 Minutes via the Falcon Fast and Light-weight task execution framework

# All-Pairs Workload 500x500 on 200 CPUs





# Mythbusting



- ~~Embarrassingly~~ Happily parallel apps are trivial to run
  - Logistical problems can be tremendous
- Loosely coupled apps do not require “supercomputers”
  - Total computational requirements can be enormous
  - Individual tasks may be tightly coupled
  - Workloads frequently involve large amounts of I/O
  - Make use of idle resources from “supercomputers” via backfilling
  - Costs to run “supercomputers” per FLOP is among the best
    - BG/P: 0.35 gigaflops/watt (**higher is better**)
    - SiCortex: 0.32 gigaflops/watt
    - BG/L: 0.23 gigaflops/watt
    - x86-based HPC systems: an order of magnitude lower
- Loosely coupled apps do not require specialized system software
- Shared file systems are good for all applications
  - They don’t scale proportionally with the compute resources
  - Data intensive applications don’t perform and scale well

# More Information



- More information: <http://people.cs.uchicago.edu/~iraicu/>
- Falkon: <http://dev.globus.org/wiki/Incubator/Falkon>
- Funding:
  - **NASA**: Ames Research Center, Graduate Student Research Program
    - Jerry C. Yan, NASA GSRP Research Advisor
  - **DOE**: Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Dept. of Energy
  - **NSF**: TeraGrid
- Check out Falkon:
  - “svn co <https://svn.globus.org/repos/falkon>”