



THE UNIVERSITY OF
CHICAGO



Argonne
NATIONAL LABORATORY

Falkon, a Fast and Light-weight task executiON framework for Clusters, Grids, and Supercomputers

Ioan Raicu

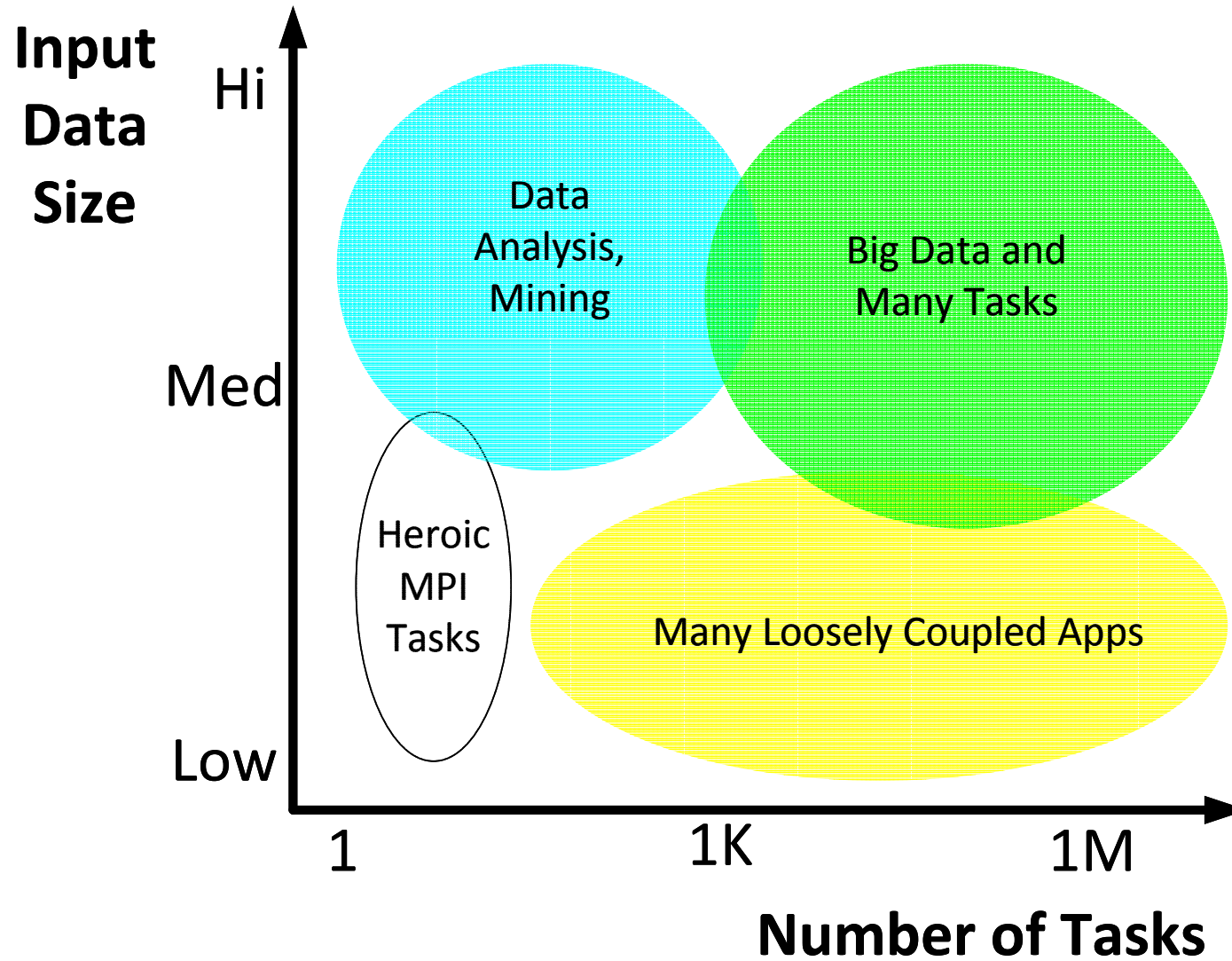
Distributed Systems Laboratory
Computer Science Department
University of Chicago

In Collaboration with:

Ian Foster, Mike Wilde, Zhao Zhang, Catalin Dumitrescu, Yong Zhao,
Pete Beckman, Kamil Iskra, Alex Szalay, Philip Little, Christopher Moretti,
Amitabh Chaudhary, Douglas Thain, Ben Clifford, plus others...

IEEE/ACM Supercomputing 2008
Argonne National Laboratory Booth
November 19th, 2008

Problem Types

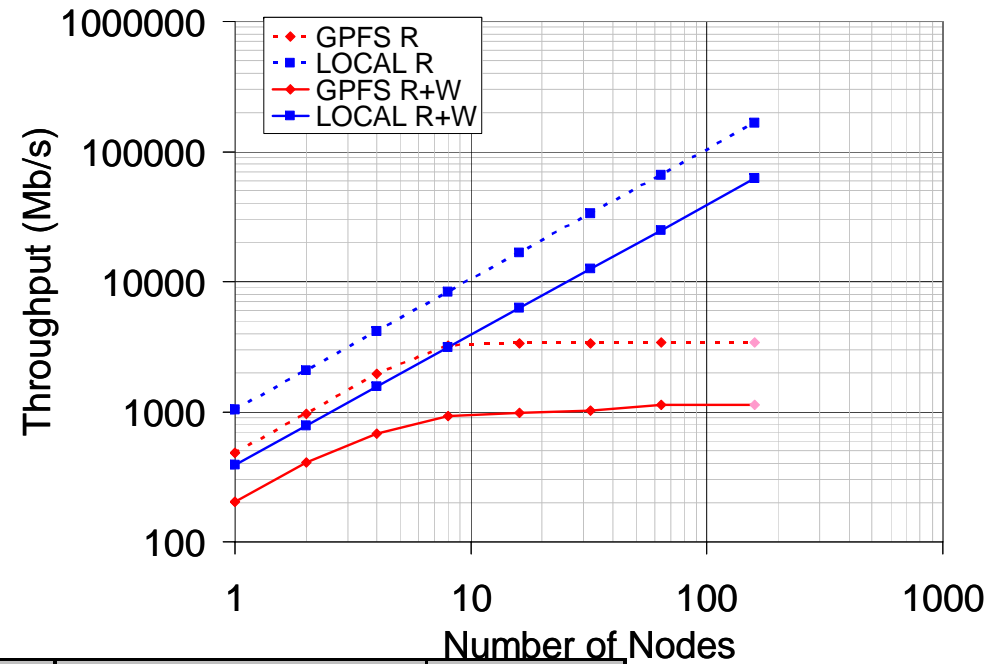
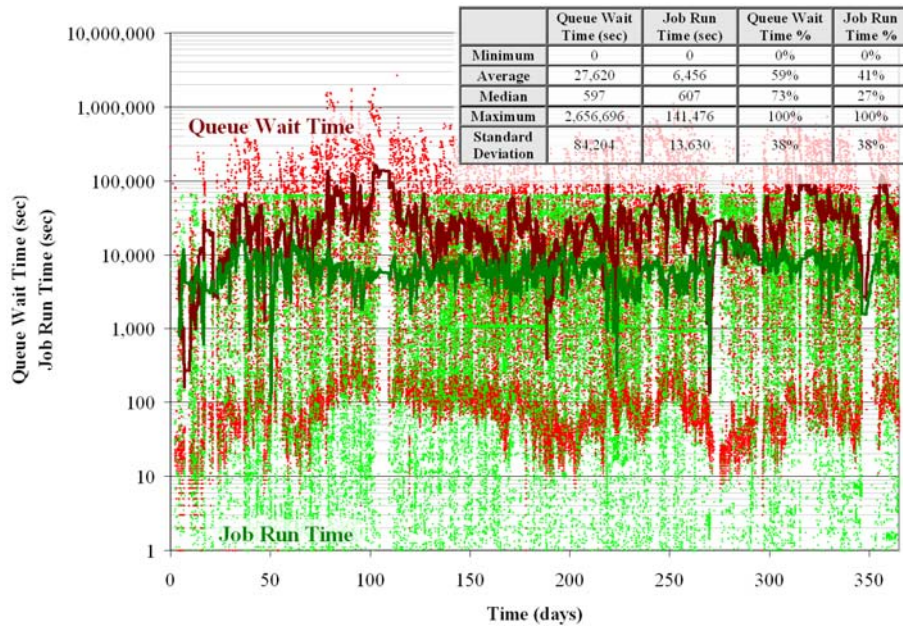


MTC: Many Task Computing



- Bridge the gap between HPC and HTC
- Loosely coupled applications with HPC orientations
- HPC comprising of multiple distinct activities, coupled via file system operations or message passing
- Emphasis on many resources over short time periods
- Tasks can be:
 - small or large, independent and dependent, uniprocessor or multiprocessor, compute-intensive or data-intensive, static or dynamic, homogeneous or heterogeneous, loosely or tightly coupled, large number of tasks, large quantity of computing, and large volumes of data...

Obstacles running MTC apps in Clusters/Grids



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

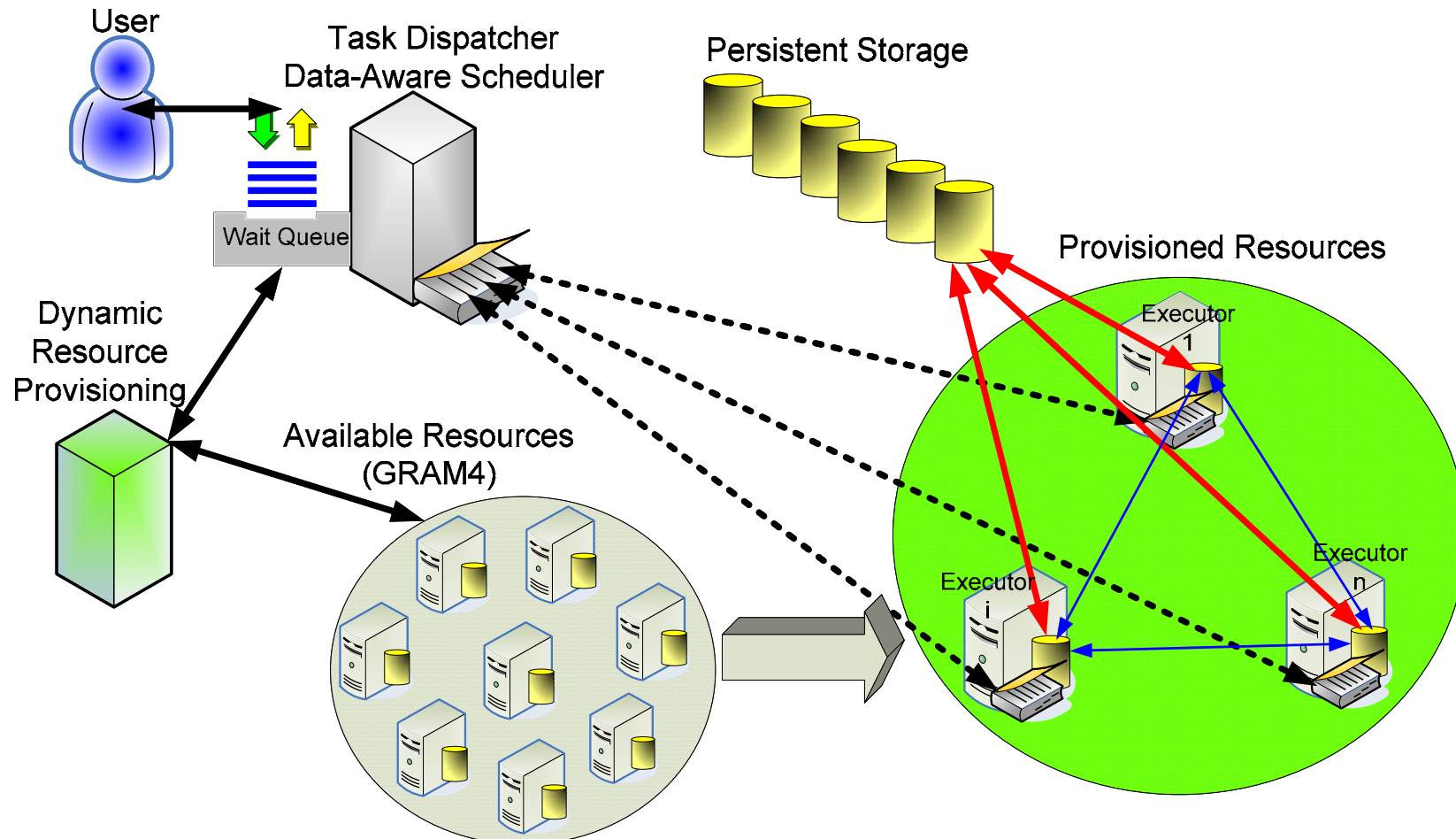
Falkon, a Fast and Light-weight task executiON framework for Clusters, Grids, and Supercomputers

Solutions



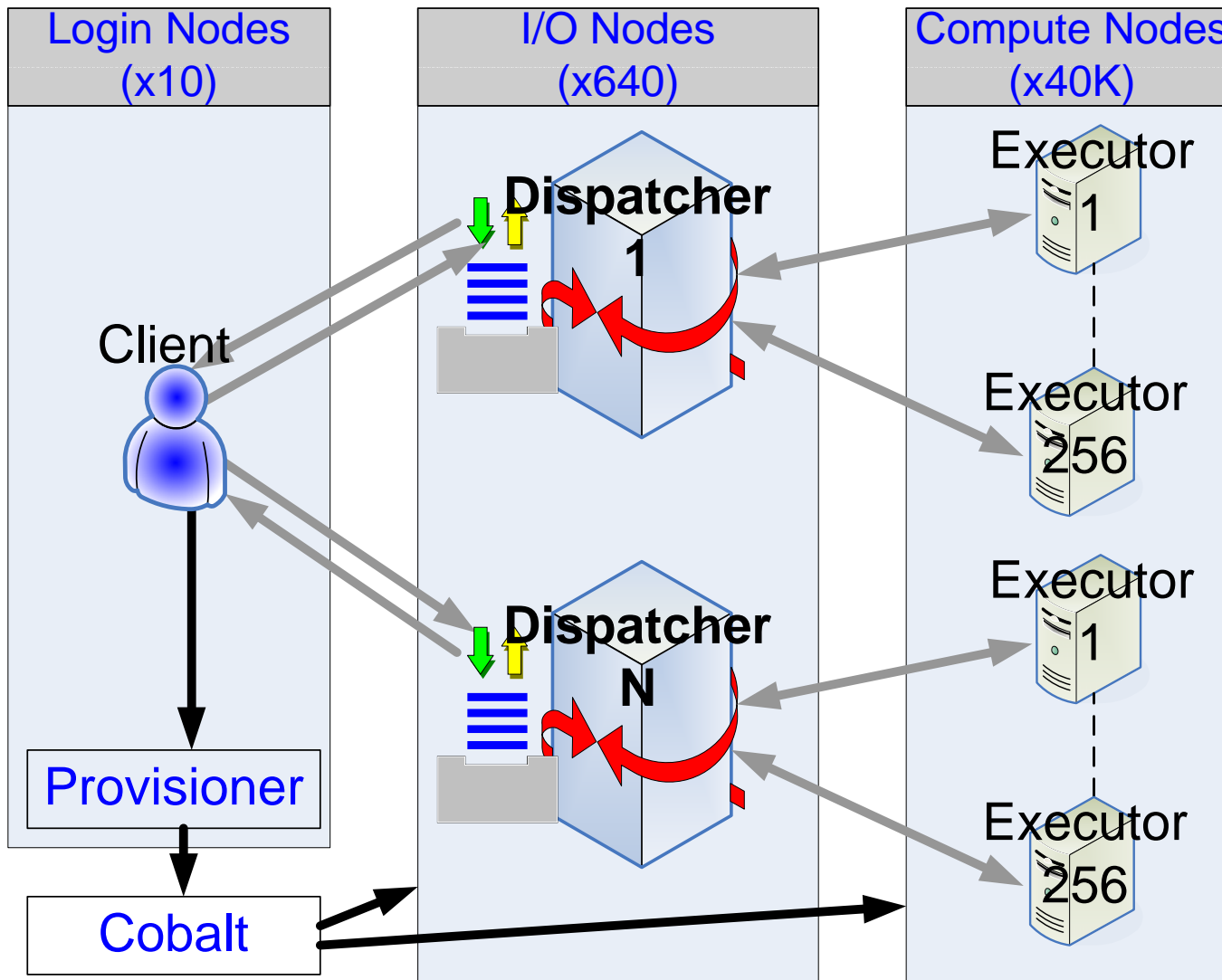
- Falkon: A Fast and Light-weight task executiON framework
 - **Goal:** enable the *rapid and efficient* execution of many independent jobs on large compute clusters
 - Combines three components:
 - A *streamlined task dispatcher*
 - *Resource provisioning* through multi-level scheduling techniques
 - *Data diffusion* and data-aware scheduling to leverage the co-located computational and storage resources
- Swift: A parallel programming system for loosely coupled applications
 - Applications cover many domains: Astronomy, astro-physics, medicine, chemistry, economics, climate modeling, data analytics

Falkon Overview

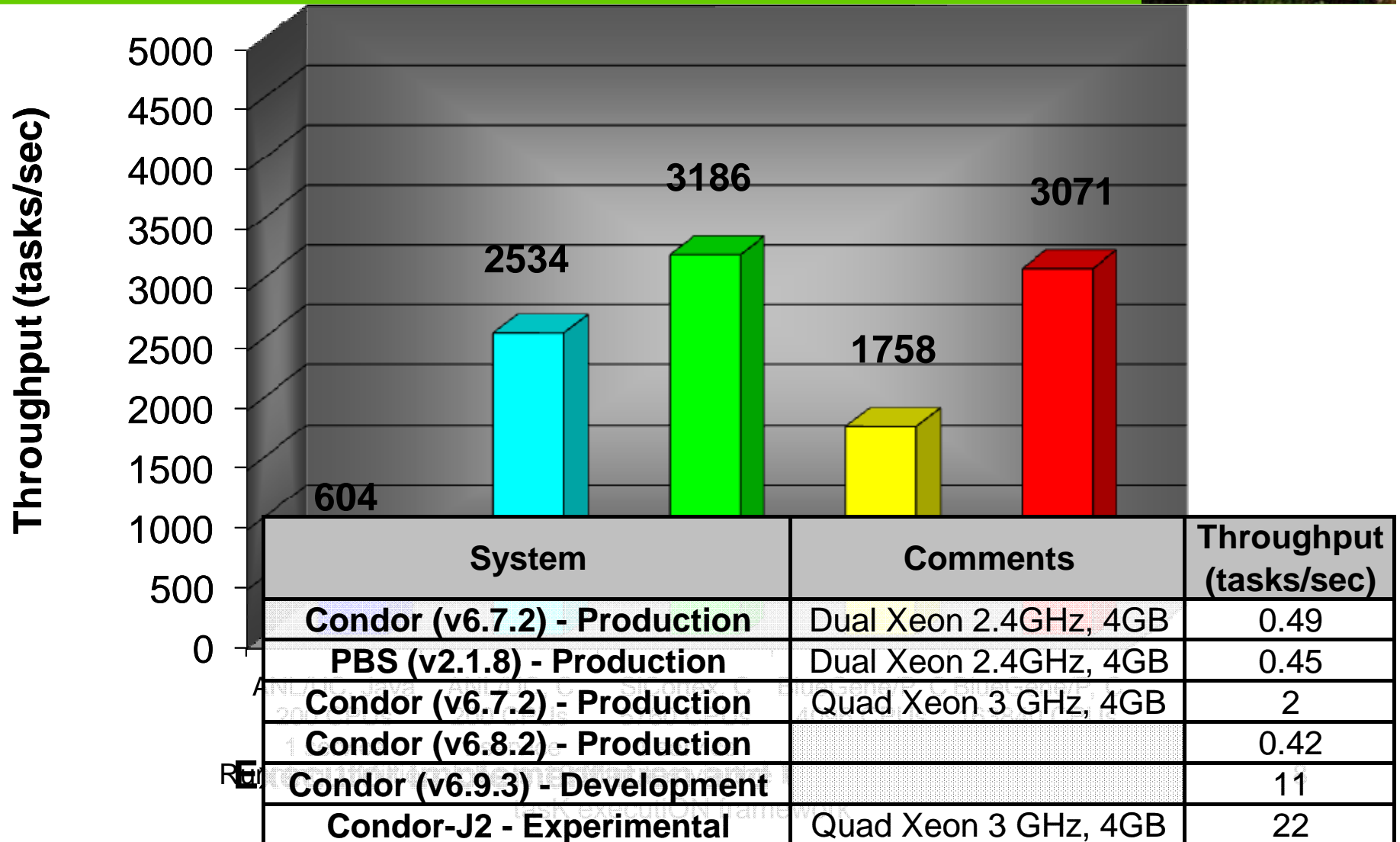


Falkon, a Fast and Light-weight task executiON framework for Clusters, Grids, and Supercomputers

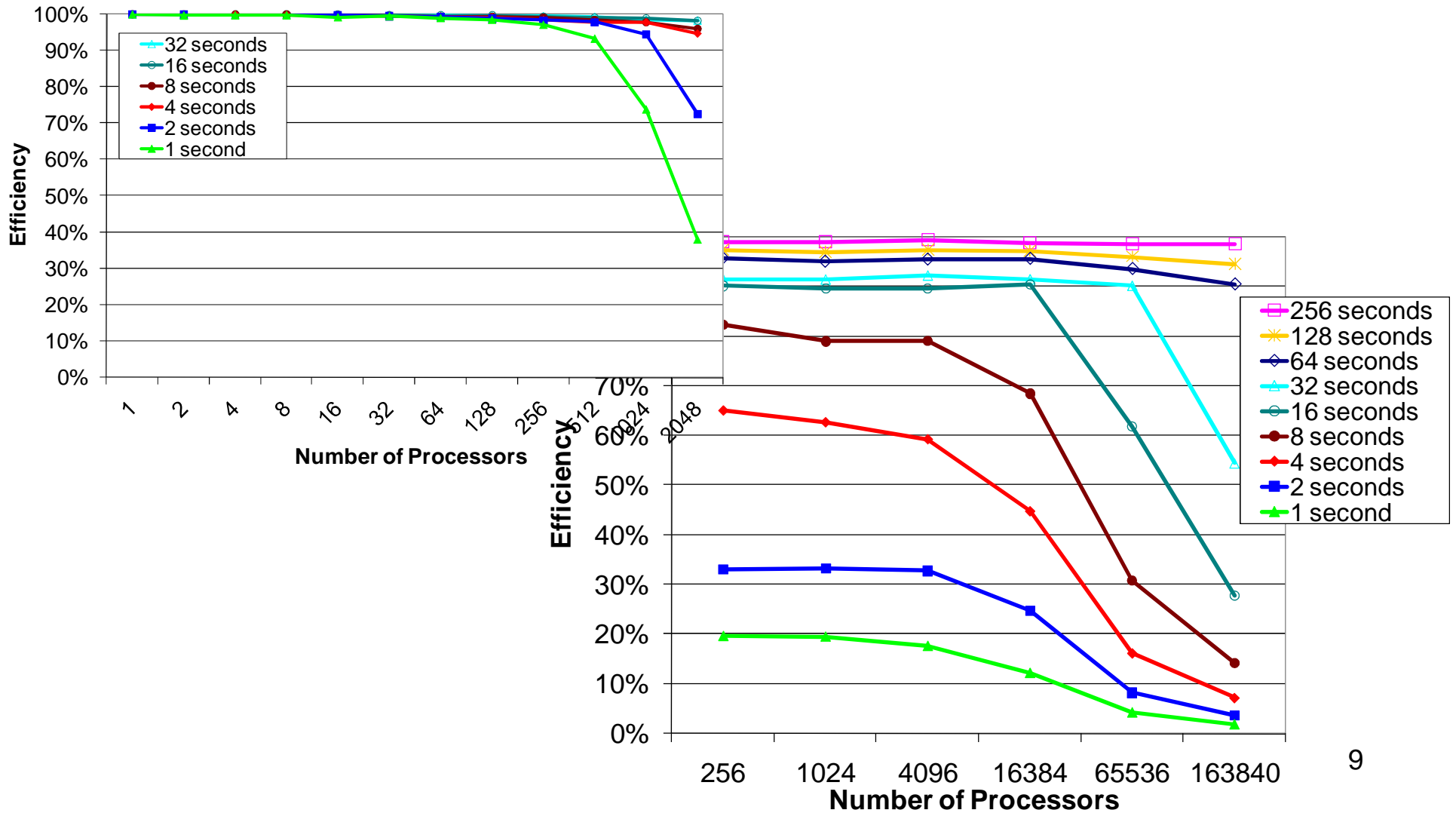
Distributed Falkon Architecture



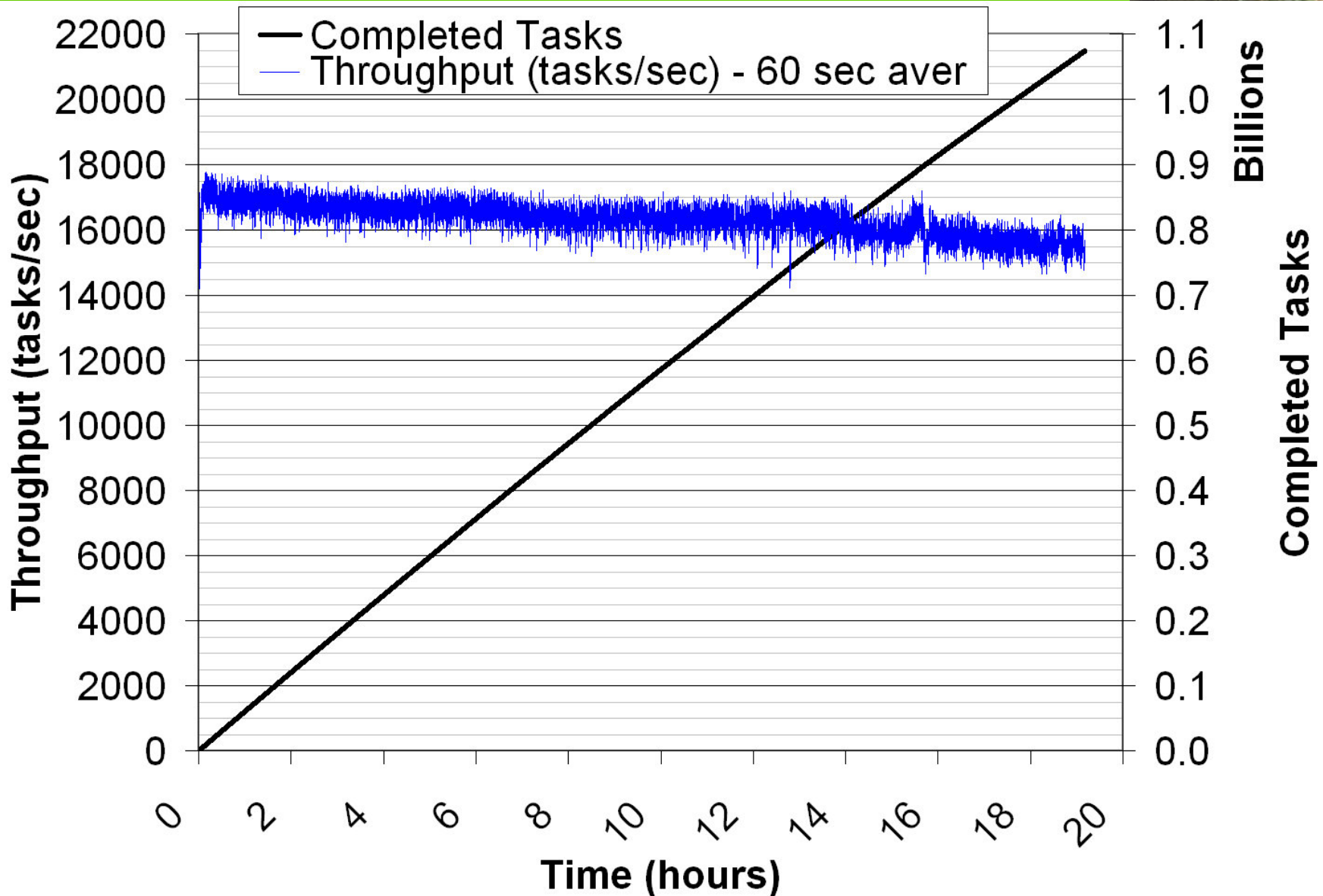
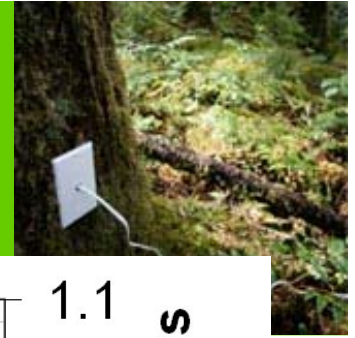
Dispatch Throughput



Efficiency



Falkon Endurance Test



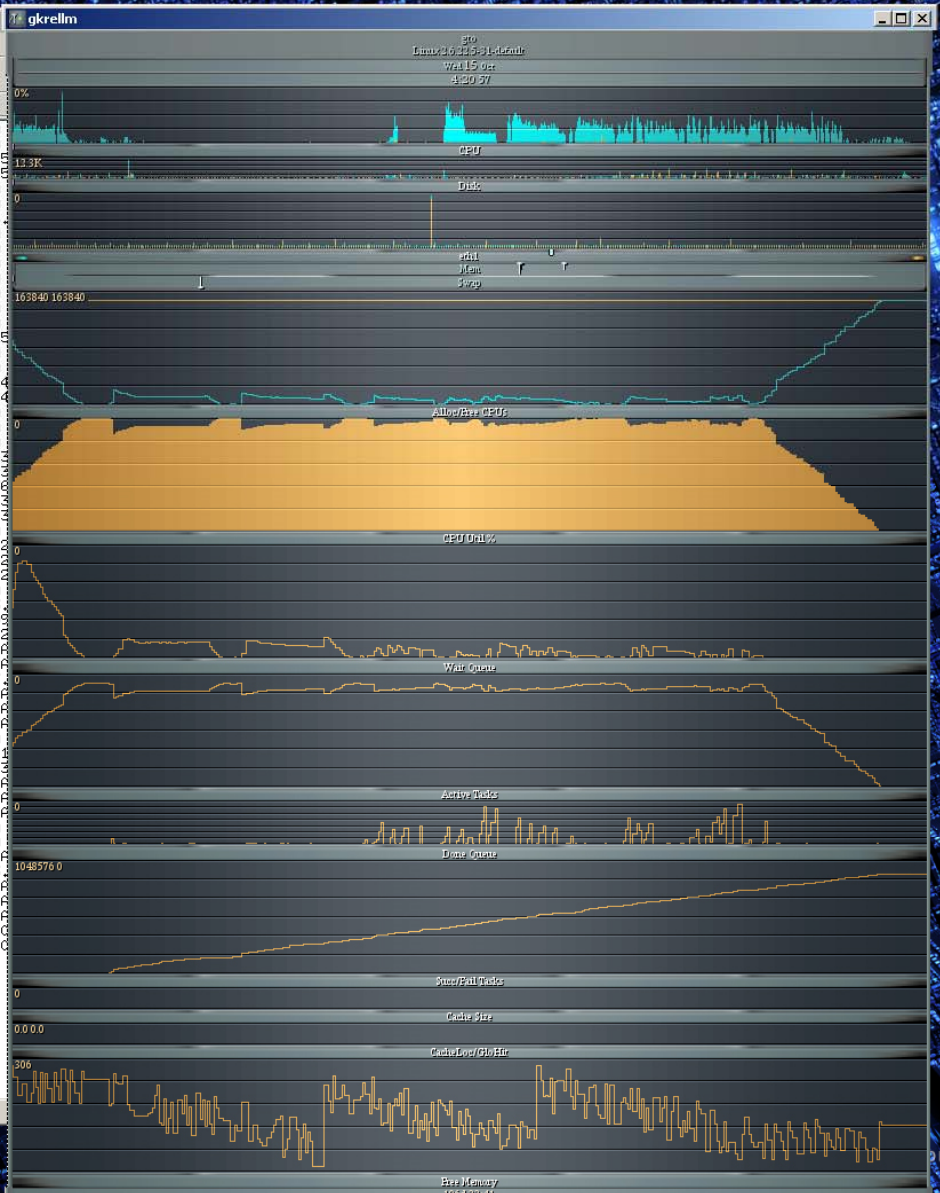
Falkon Monitoring

```
gto.ci.uchicago.edu (1) - SecureCRT
File Edit View Options Transfer Script Tools Help

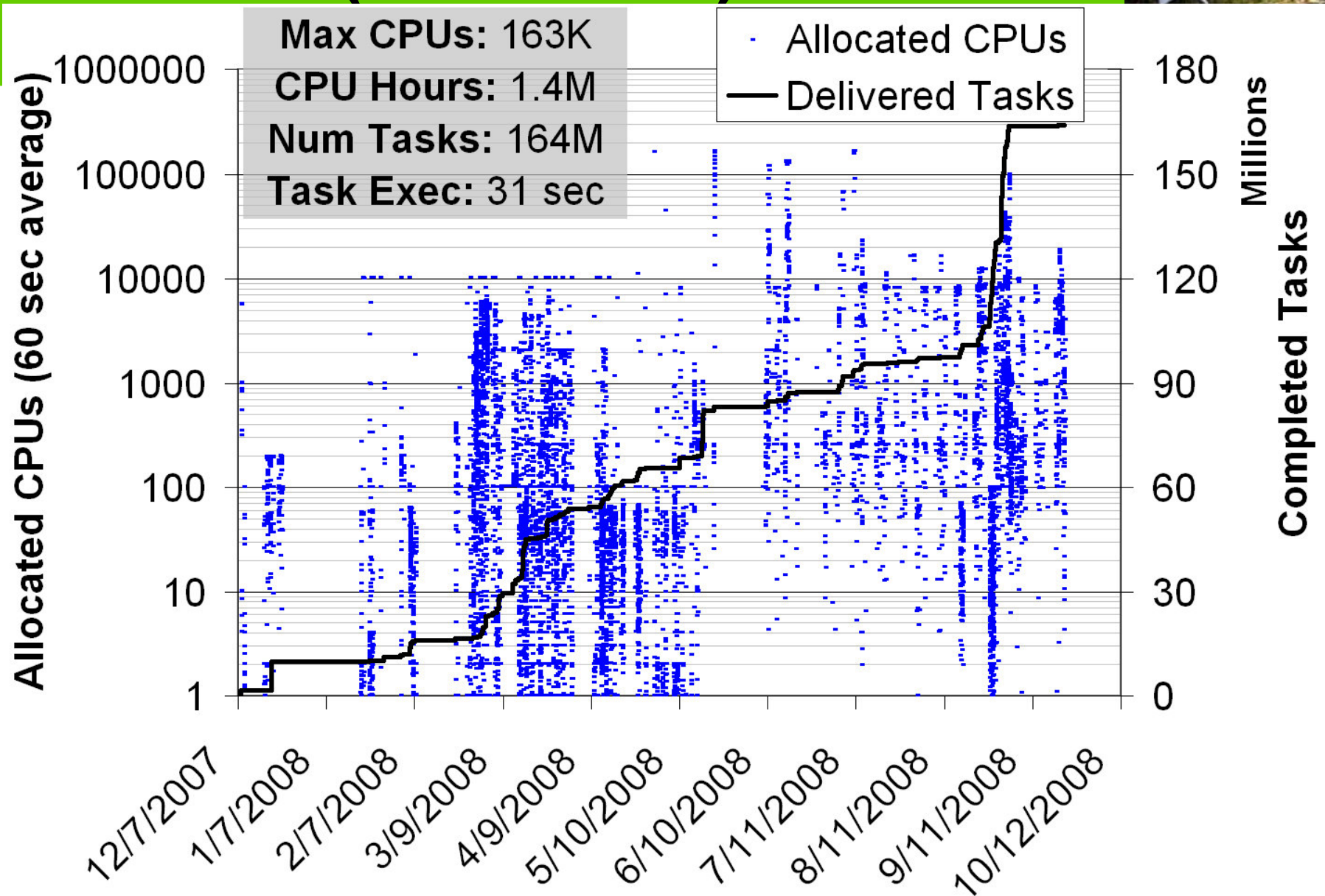
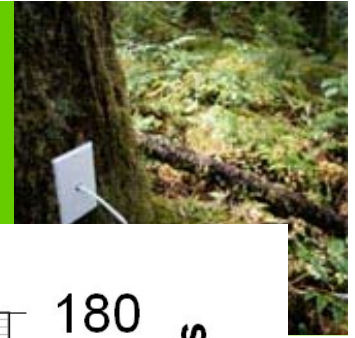
gto.ci.uchicago.edu | gto.ci.uchicago.edu (1) | gto.ci.uchicago.edu (3) | gto.ci.uchicago.edu (2) | gto.ci.uchicago.edu (5) | gto.ci.uchicago.edu (4)

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 89.89 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 3540.84 aver_tp 2692.23 stdev_tp 3111.472 ETA
409,021 tasks+ 937283 tasks- 0 tasks-> 1048576 completed 89.36 tasks_tp 3439.24 aver_tp 2691.43 stdev_tp 3106.976 ETA
410,025 tasks+ 939976 tasks- 0 tasks-> 1048576 completed 89.57 tasks_tp 3122.51 aver_tp 2689.84 stdev_tp 3102.621 ETA
411,029 tasks+ 942688 tasks- 0 tasks-> 1048576 completed 89.79 tasks_tp 3279.85 aver_tp 2688.65 stdev_tp 3098.212 ETA
412,033 tasks+ 945416 tasks- 0 tasks-> 1048576 completed 90.0 tasks_tp 2219.13 aver_tp 2687.3 stdev_tp 3093.948 ETA
413,038 tasks+ 948166 tasks- 0 tasks-> 1048576 completed 90.21 tasks_tp 3171.31 aver_tp 2685.81 stdev_tp 3089.523 ETA
414,042 tasks+ 949926 tasks- 0 tasks-> 1048576 completed 90.42 tasks_tp 3234.06 aver_tp 2684.52 stdev_tp 3085.188 ETA
415,046 tasks+ 951703 tasks- 0 tasks-> 1048576 completed 90.62 tasks_tp 3047.81 aver_tp 2682.7 stdev_tp 3080.963 ETA
416,050 tasks+ 953505 tasks- 0 tasks-> 1048576 completed 90.81 tasks_tp 2144.42 aver_tp 2681.17 stdev_tp 3076.707 ETA
417,054 tasks+ 954951 tasks- 0 tasks-> 1048576 completed 91.0 tasks_tp 2214.14 aver_tp 2679.84 stdev_tp 3072.434 ETA
418,062 tasks+ 956445 tasks- 0 tasks-> 1048576 completed 91.18 tasks_tp 2067.46 aver_tp 2678.11 stdev_tp 3068.251 ETA
419,071 tasks+ 958742 tasks- 0 tasks-> 1048576 completed 91.43 tasks_tp 2090.36 aver_tp 2676.42 stdev_tp 3064.079 ETA
420,079 tasks+ 960450 tasks- 0 tasks-> 1048576 completed 91.6 tasks_tp 1724.31 aver_tp 2673.73 stdev_tp 3060.176 ETA
421,087 tasks+ 962605 tasks- 0 tasks-> 1048576 completed 91.8 tasks_tp 2108.13 aver_tp 2672.15 stdev_tp 3056.022 ETA
422,095 tasks+ 964675 tasks- 0 tasks-> 1048576 completed 92.0 tasks_tp 2075.4 aver_tp 2670.47 stdev_tp 3051.902 ETA
423,103 tasks+ 966960 tasks- 0 tasks-> 1048576 completed 92.12 tasks_tp 1248.02 aver_tp 2666.5 stdev_tp 3048.561 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+ 976213 tasks- 0 tasks-> 1048576 completed 93.29 tasks_tp 3722.22 aver_tp 2695.43 stdev_tp 3065.644 ETA
426,128 tasks+ 980343 tasks- 0 tasks-> 1048576 completed 93.48 tasks_tp 2011.9 aver_tp 2683.57 stdev_tp 3051.614 ETA
427,136 tasks+ 982449 tasks- 0 tasks-> 1048576 completed 93.87 aver_tp 2682.19 stdev_tp 3047.908 ETA
428,144 tasks+ 983491 tasks- 0 tasks-> 1048576 completed 94.31 aver_tp 2677.45 stdev_tp 3044.643 ETA
429,152 tasks+ 98763 tasks- 0 tasks-> 1048576 completed 94.76 aver_tp 2691.51 stdev_tp 3041.641 ETA
430,16 tasks+ 995260 tasks- 0 tasks-> 1048576 completed 94.26 aver_tp 2694.57 stdev_tp 3048.1 ETA
431,168 tasks+ 995260 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 379.97 aver_tp 2691.53 stdev_tp 3041.282 ETA
434,192 tasks+ 1000723 tasks- 0 tasks-> 1048576 completed 95.99 tasks_tp 191.53 stdev_tp 3040.335 ETA
435,201 tasks+ 1000723 tasks- 0 tasks-> 1048576 completed 96.47 tasks_tp 268.26 aver_tp 2694.44 stdev_tp 3041.177 ETA
436,209 tasks+ 1011812 tasks- 0 tasks-> 1048576 completed 96.49 tasks_tp 1927.58 aver_tp 2695.38 stdev_tp 3031.597 ETA
437,217 tasks+ 1011812 tasks- 0 tasks-> 1048576 completed 96.84 tasks_tp 3585.56 aver_tp 2690.71 stdev_tp 3027.853 ETA
438,225 tasks+ 1013995 tasks- 0 tasks-> 1048576 completed 97.27 tasks_tp 4850.6 aver_tp 2695.0 stdev_tp 3025.337 ETA
439,233 tasks+ 1013995 tasks- 0 tasks-> 1048576 completed 97.77 tasks_tp 102.898 ETA
440,241 tasks+ 1033223 tasks- 0 tasks-> 1048576 completed 98.25 tasks_tp 4472.22 aver_tp 2694.33 stdev_tp 3016.682 ETA
441,249 tasks+ 1032655 tasks- 0 tasks-> 1048576 completed 98.47 tasks_tp 2319.64 aver_tp 2693.04 stdev_tp 3012.127 ETA
442,257 tasks+ 1032655 tasks- 0 tasks-> 1048576 completed 98.83 tasks_tp 3662.7 aver_tp 2695.53 stdev_tp 3009.572 ETA
443,265 tasks+ 1033253 tasks- 0 tasks-> 1048576 completed 99.17 tasks_tp 2864.48 aver_tp 2695.24 stdev_tp 3004.628 ETA
444,273 tasks+ 1033253 tasks- 0 tasks-> 1048576 completed 99.3 tasks_tp 1929.56 aver_tp 2693.24 stdev_tp 3000.948 ETA
445,281 tasks+ 1046203 tasks- 0 tasks-> 1048576 completed 99.3 tasks_tp 2265.87 aver_tp 2696.15 stdev_tp 2995.499 ETA
446,289 tasks+ 1046203 tasks- 0 tasks-> 1048576 completed 99.77 tasks_tp 2688.16 aver_tp 2696.15 stdev_tp 2991.596 ETA
447,297 tasks+ 1046203 tasks- 0 tasks-> 1048576 completed 100.0 tasks_tp 2354.17 aver_tp 2685.29 stdev_tp 2987.766 ETA
448,305 tasks+ 1048576 tasks- 0 tasks-> 1048576 completed 100.0 tasks_tp 0.0 aver_tp 2678.35 stdev_tp 2987.016 ETA
449,313 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>
```

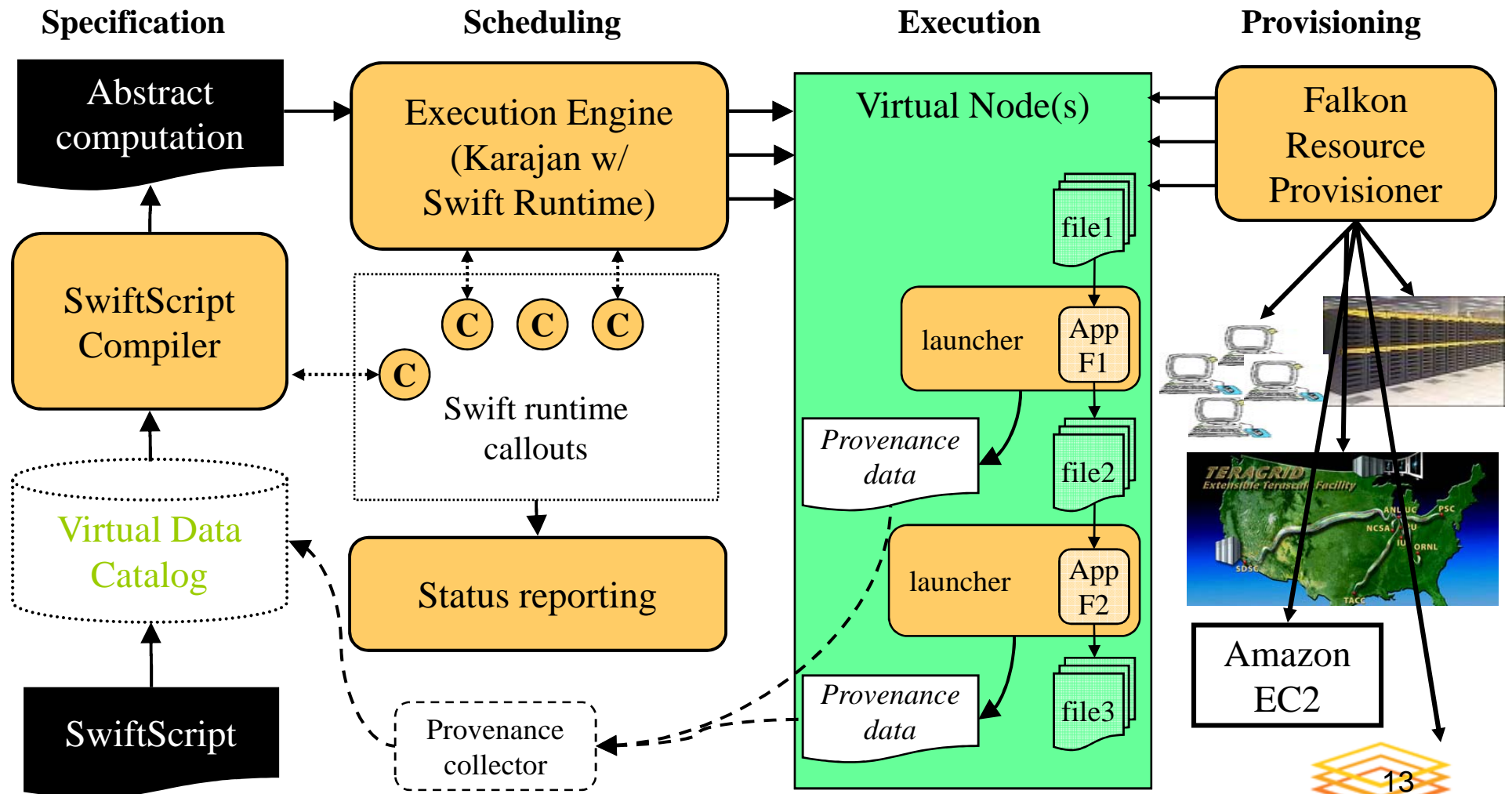
- Workload
- 160K CPUs
- 1M tasks
- 60 sec per task
- 17.5K CPU hours in 7.5 min
- Throughput: 2312 tasks/sec
- 85% efficiency



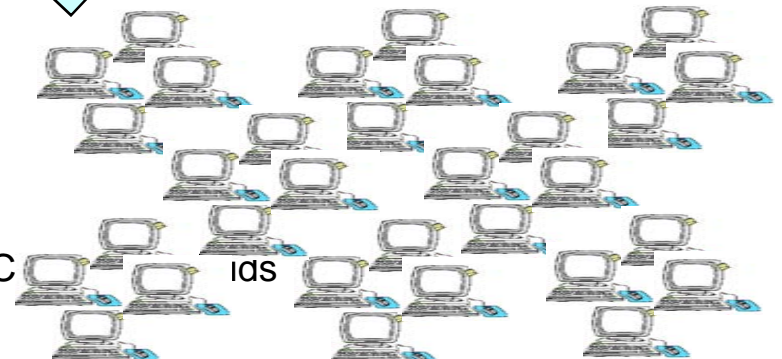
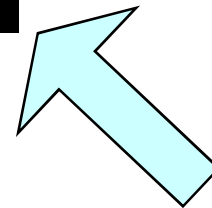
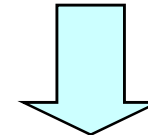
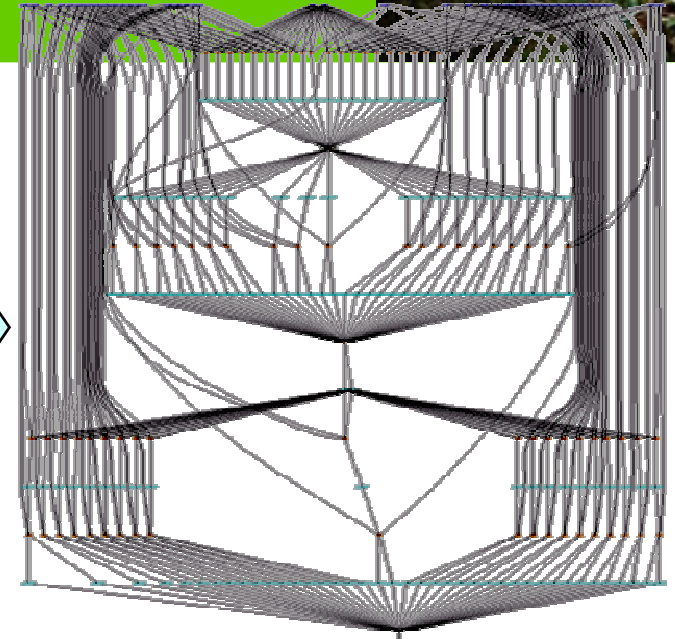
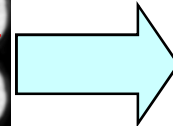
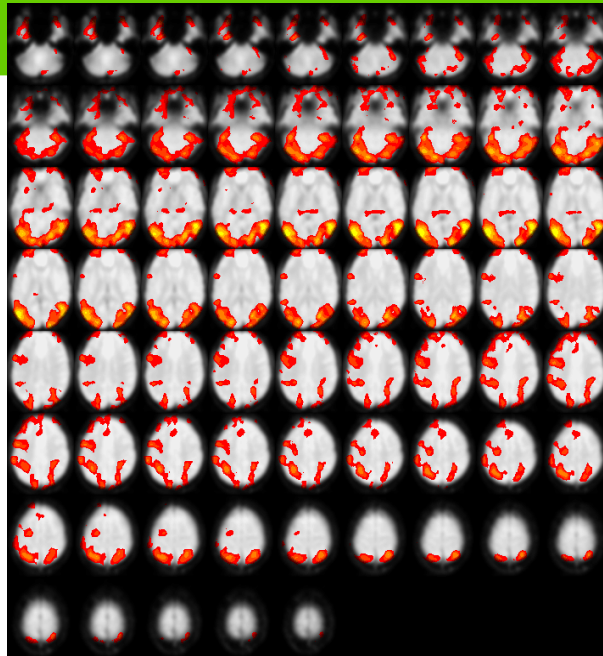
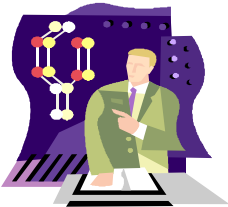
Falkon Activity History (10 months)



Swift Architecture



Functional MRI (fMRI)



ment in C

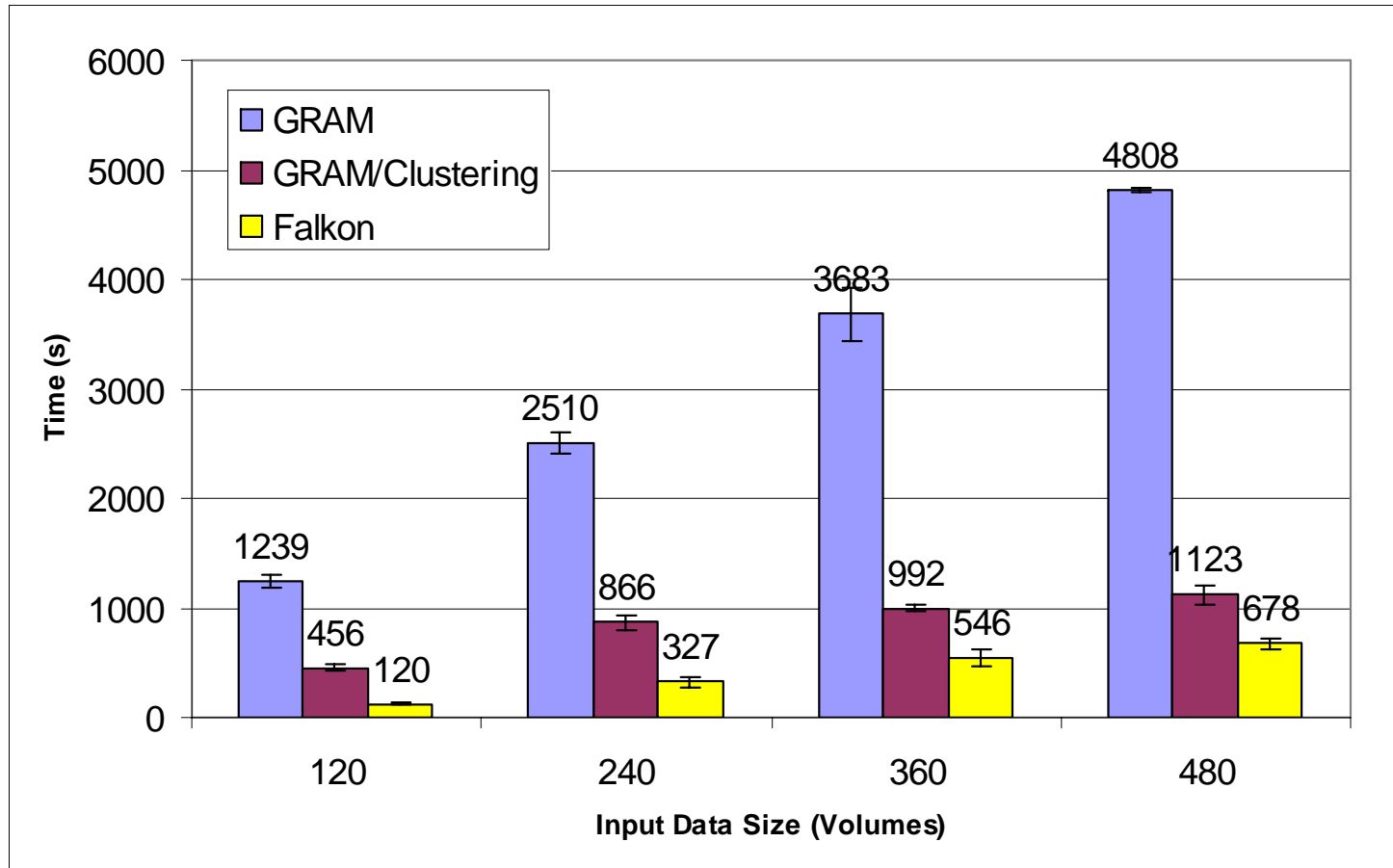
ids

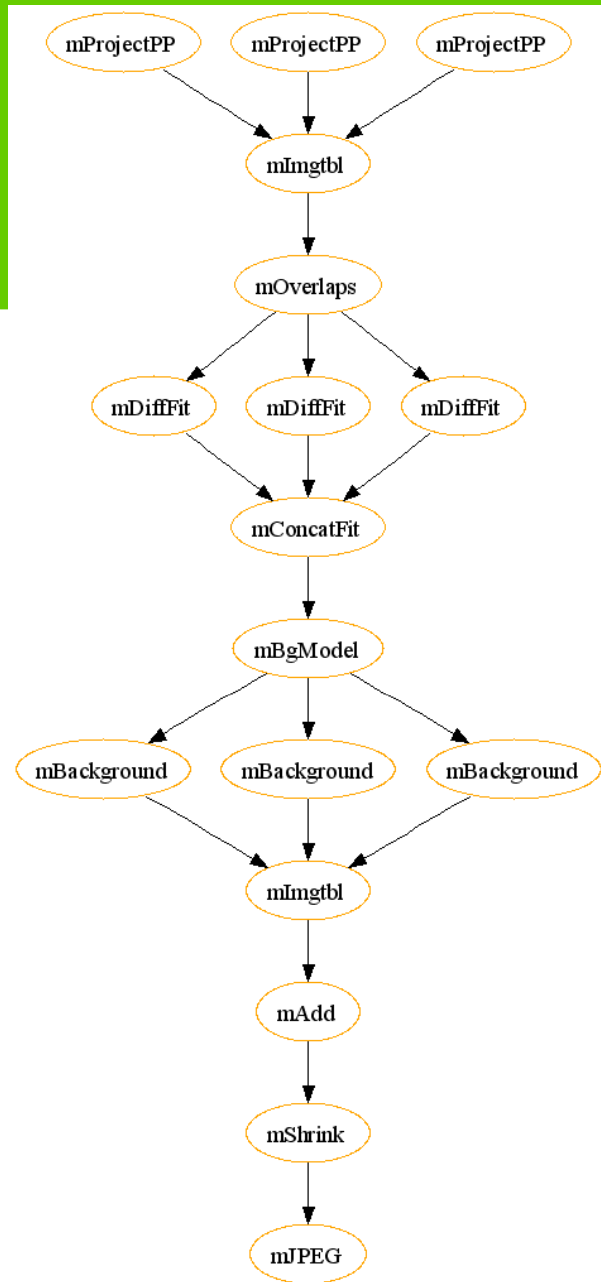
- Wide range of analyses
 - Testing, interactive analysis, production runs
 - Data mining
 - Parameter studies

Completed Milestones: fMRI Application

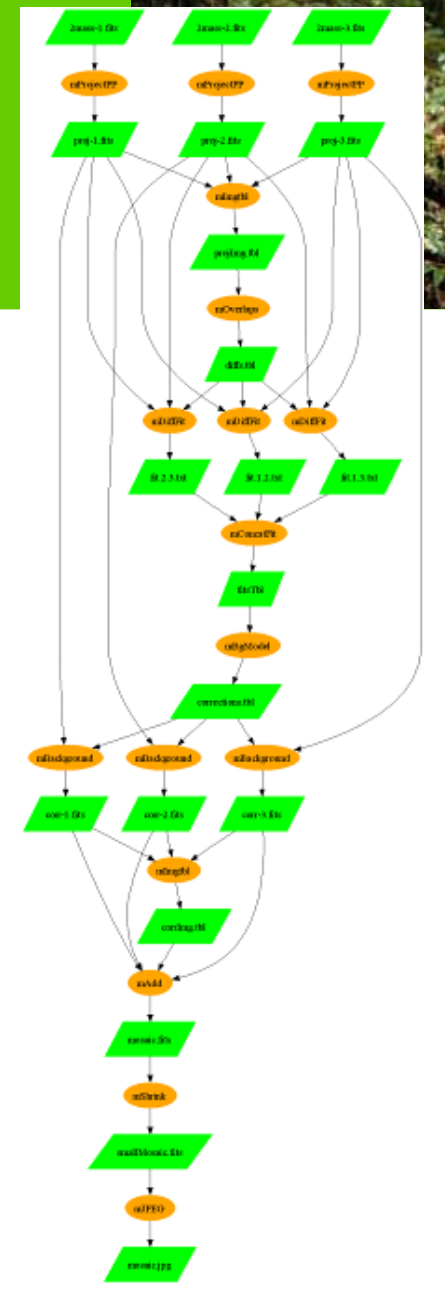


- GRAM vs. Falcon: 85%~90% lower run time
- GRAM/Clustering vs. Falcon: 40%~74% lower run time





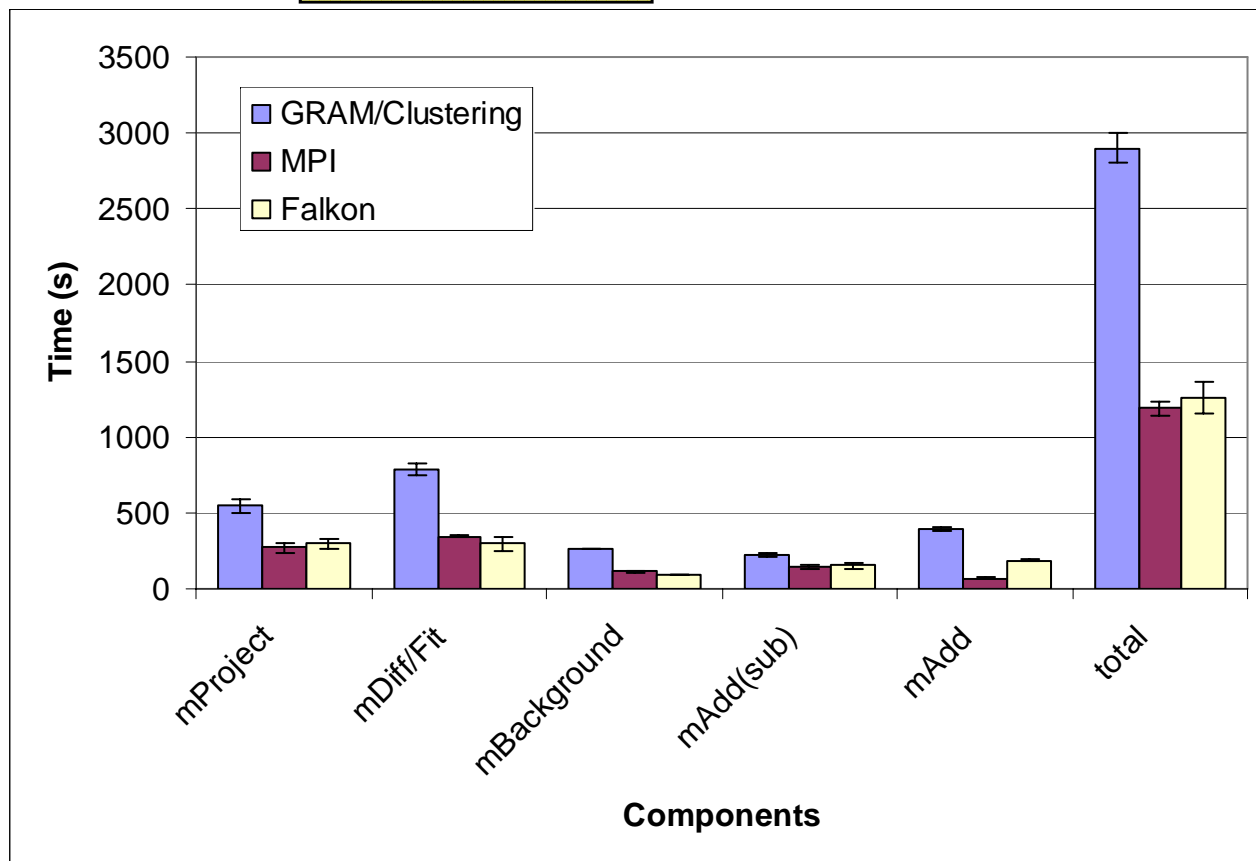
B. Berriman, J. Good (Caltech)
 J. Jacob, D. Katz (JPL)



Completed Milestones: Montage Application



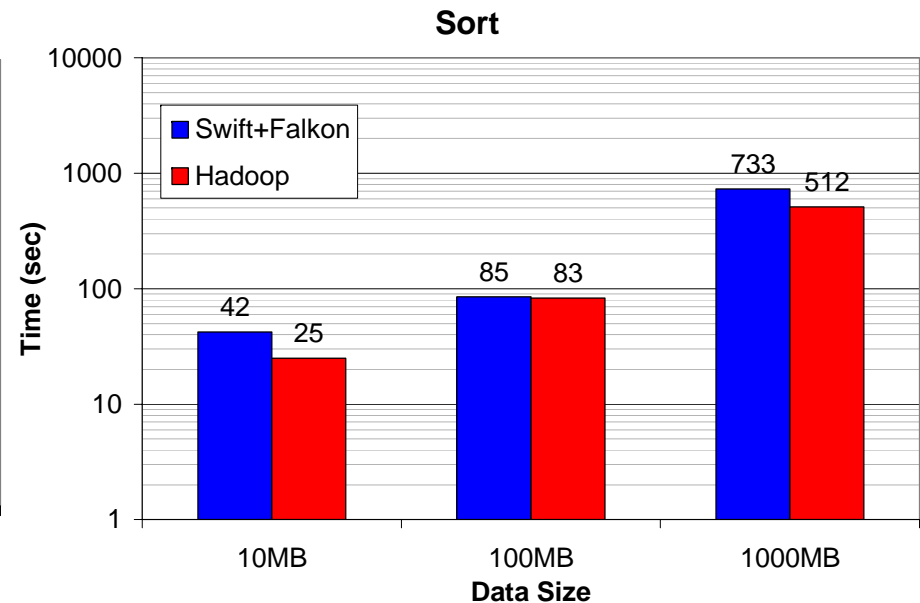
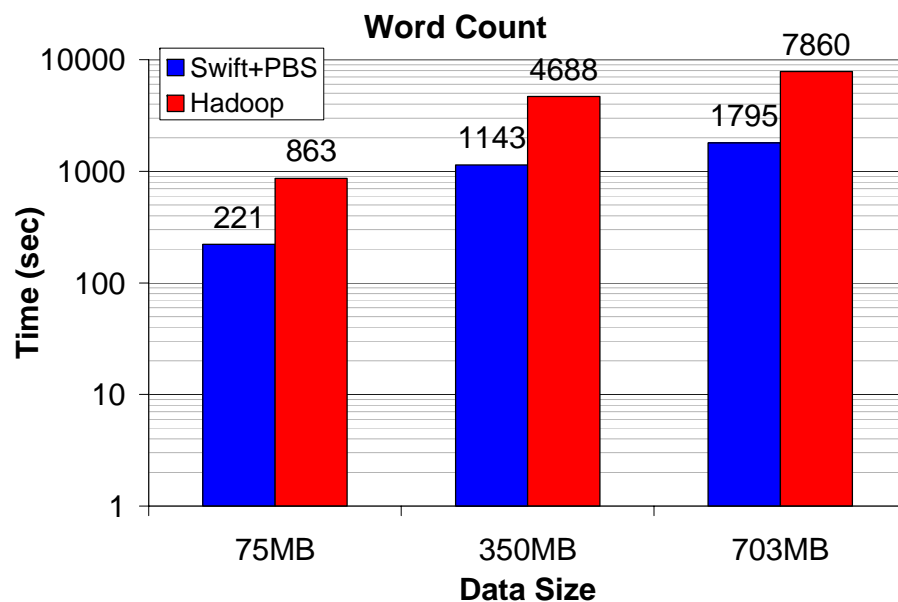
- GRAM/Clustering vs. Falcon: **57%** lower application run time
- MPI* vs. Falcon: **4%** higher application run time
- * MPI should be **lower bound**



Hadoop vs. Swift



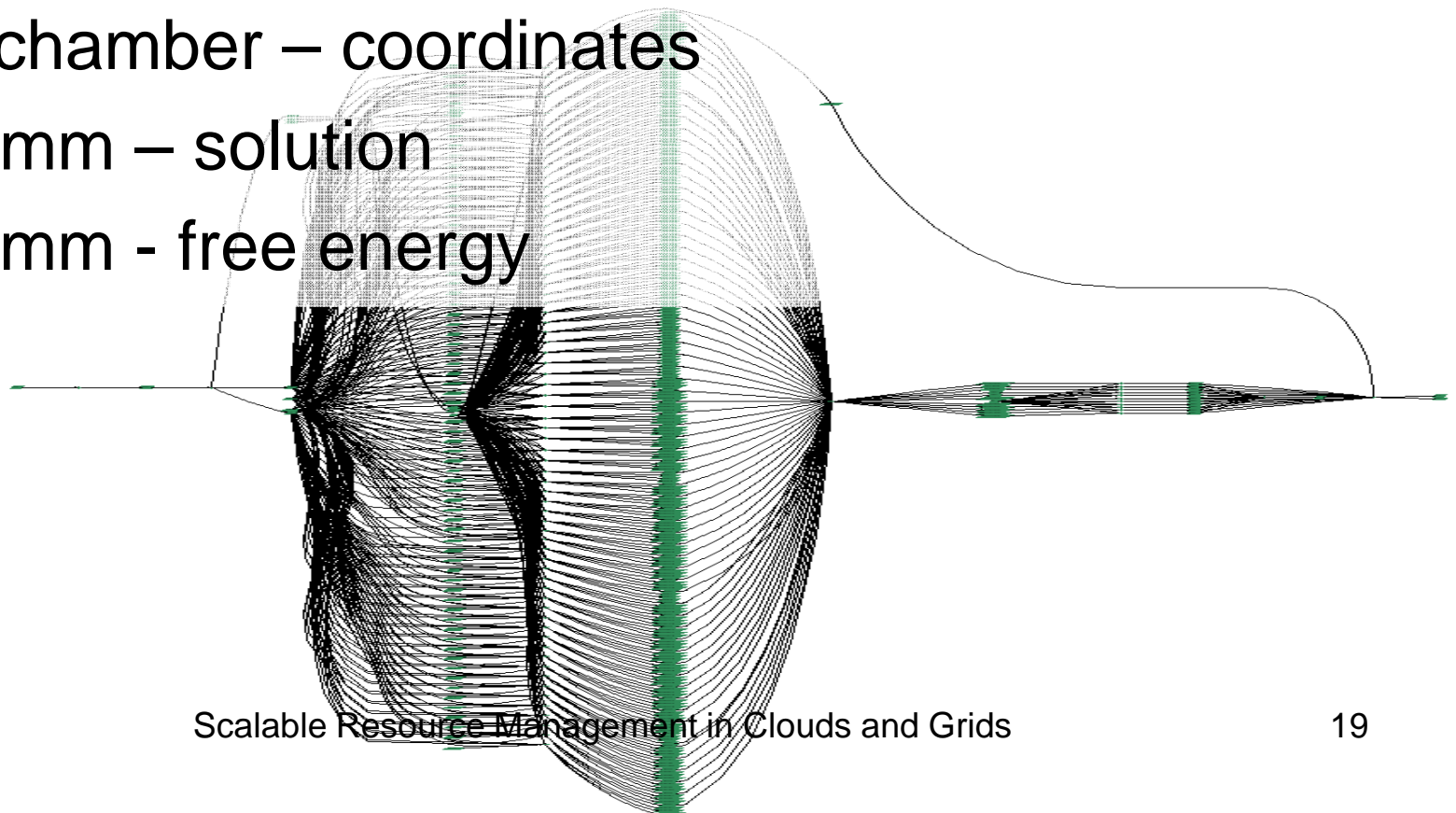
- Classic benchmarks for MapReduce
 - Word Count
 - Sort
- Swift performs similar or better than Hadoop (on 32 processors)



Molecular Dynamics

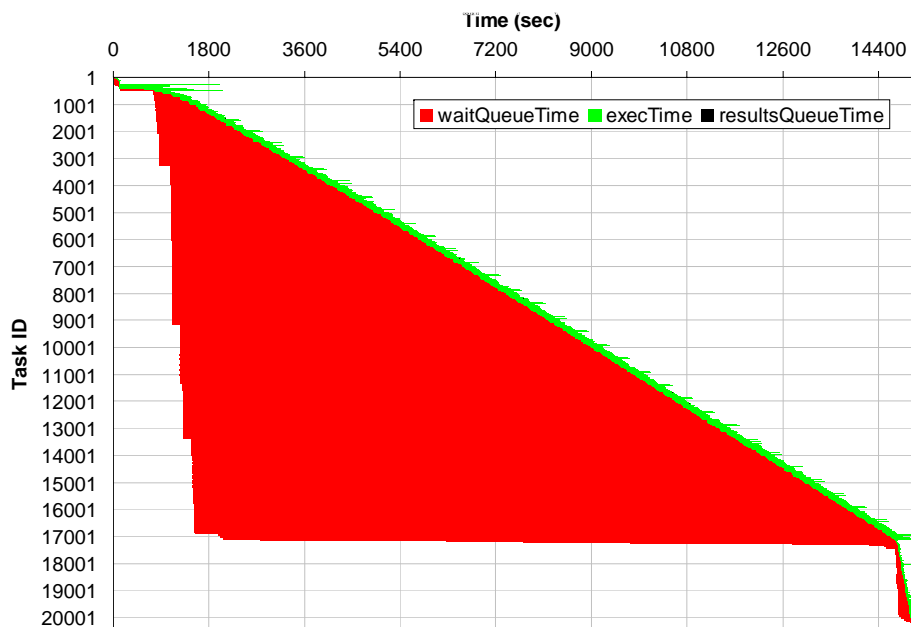


- Determination of free energies in aqueous solution
 - Antechamber – coordinates
 - Charmm – solution
 - Charmm - free energy

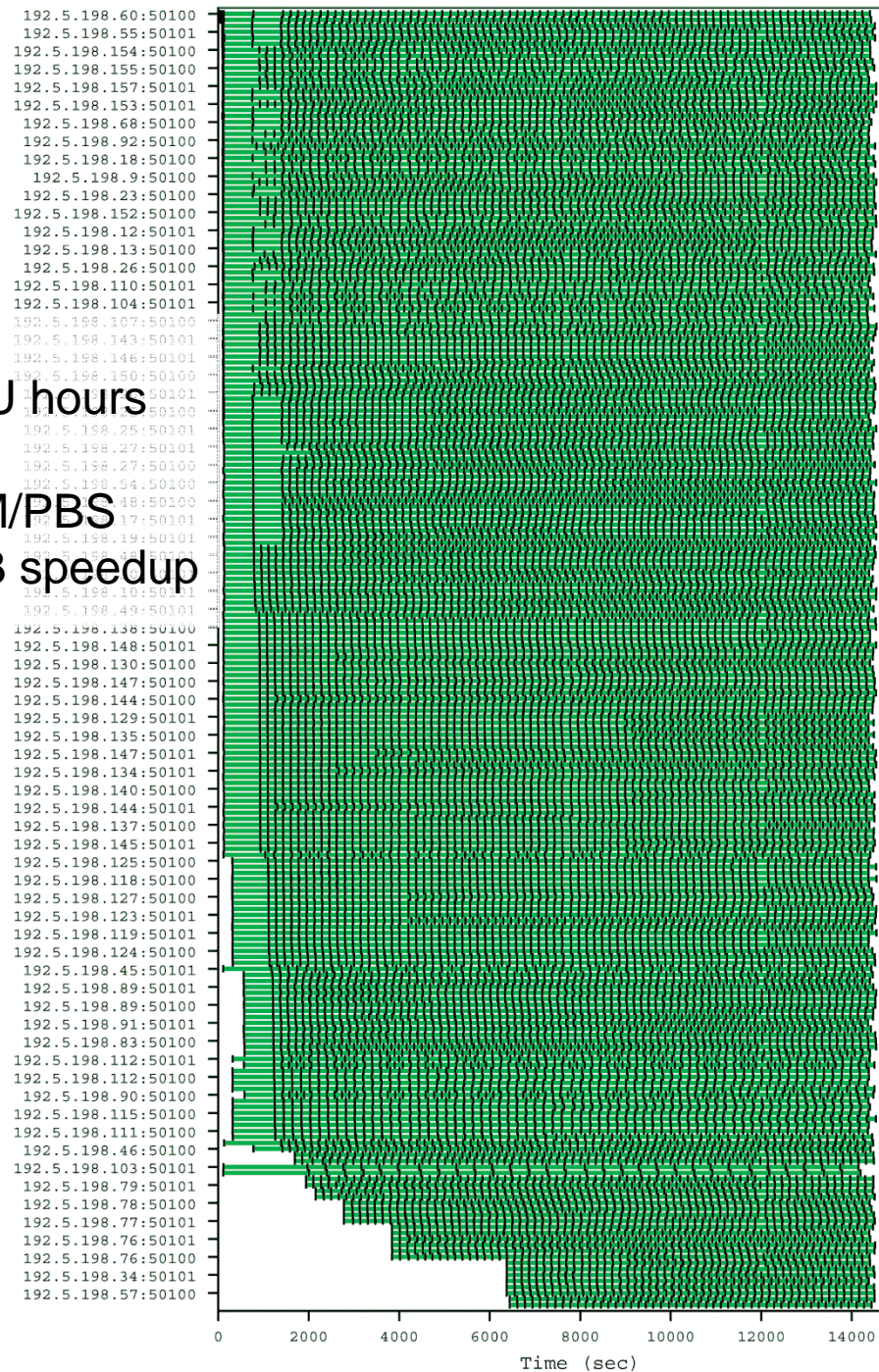


MolDyn Application

- 244 molecules → 20497 jobs
- 15091 seconds on 216 CPUs → 867.1 CPU hours
- Efficiency: **99.8%**
- Speedup: 206.9x → 8.2x faster than GRAM/PBS
- 50 molecules w/ GRAM (4201 jobs) → 25.3 speedup



Scalable Resource Manage

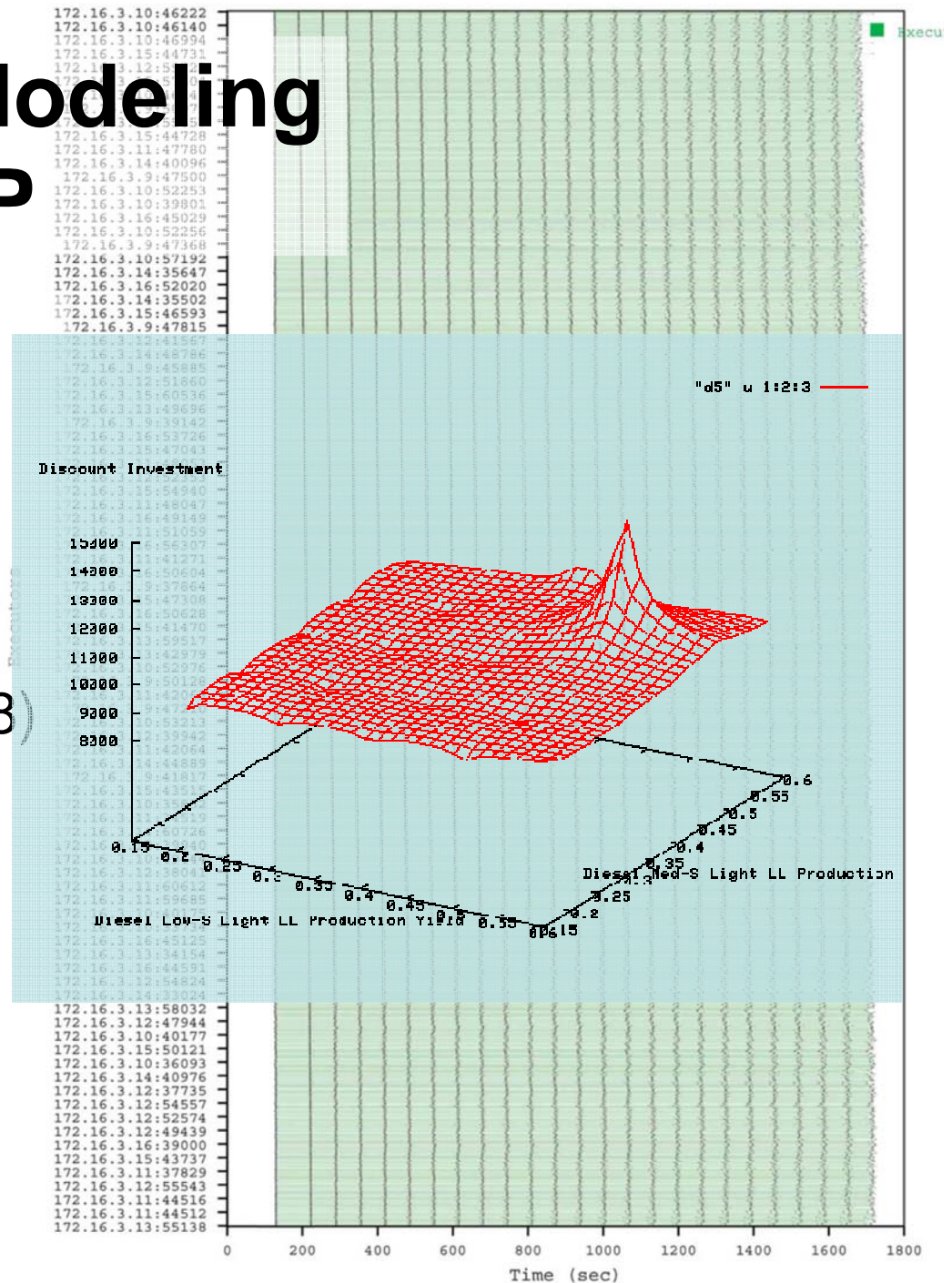


MARS Economic Modeling on IBM BG/P

- CPU Cores: 2048
- Tasks: 49152
- Micro-tasks: 7077888
- Elapsed time: 1601 secs
- CPU Hours: 894
- Speedup: 1993X (ideal 2048)
- Efficiency: 97.3%



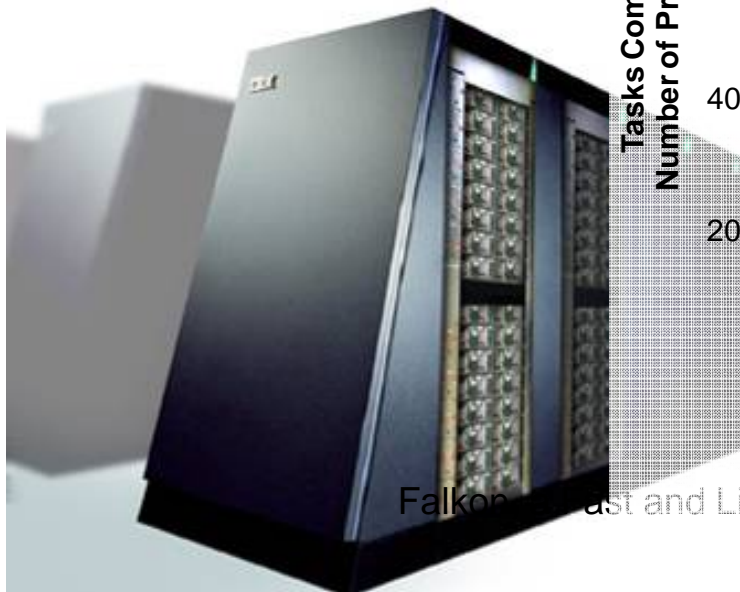
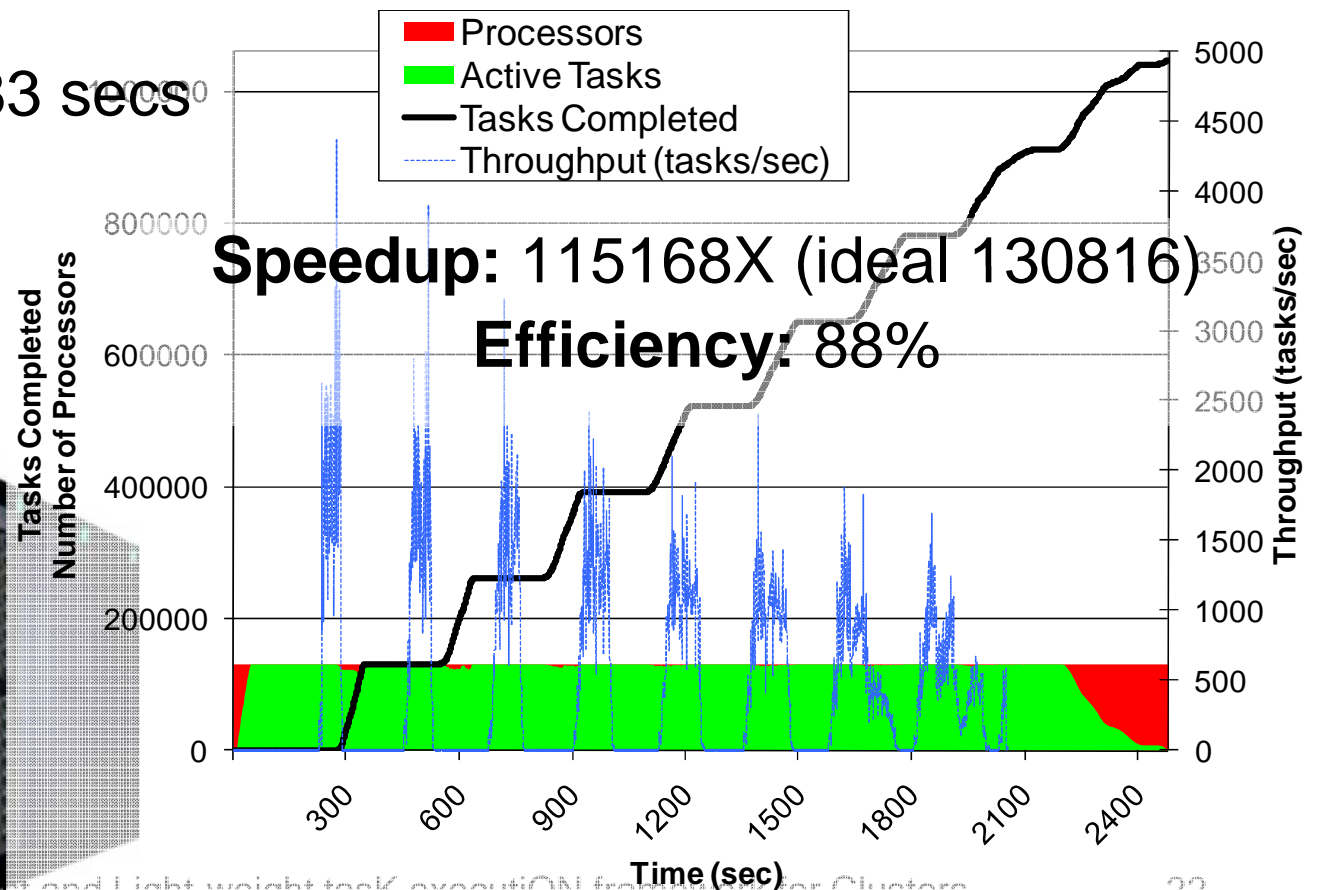
ght-weic
Grids, &



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



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



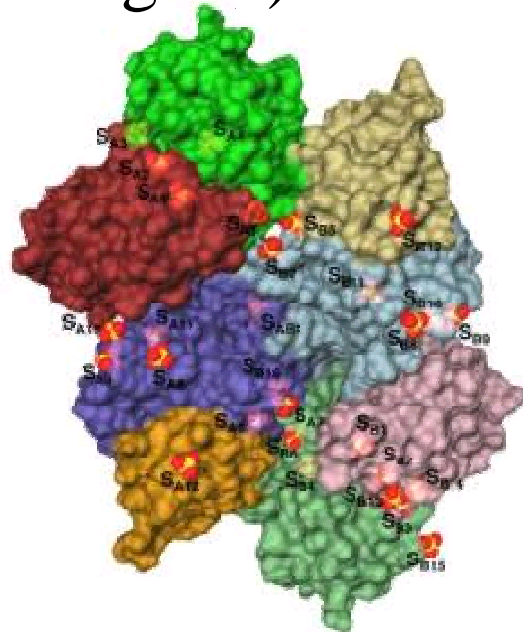
Falcon

Fast and Light-weight task execution framework for Clusters, Grids, and Supercomputers

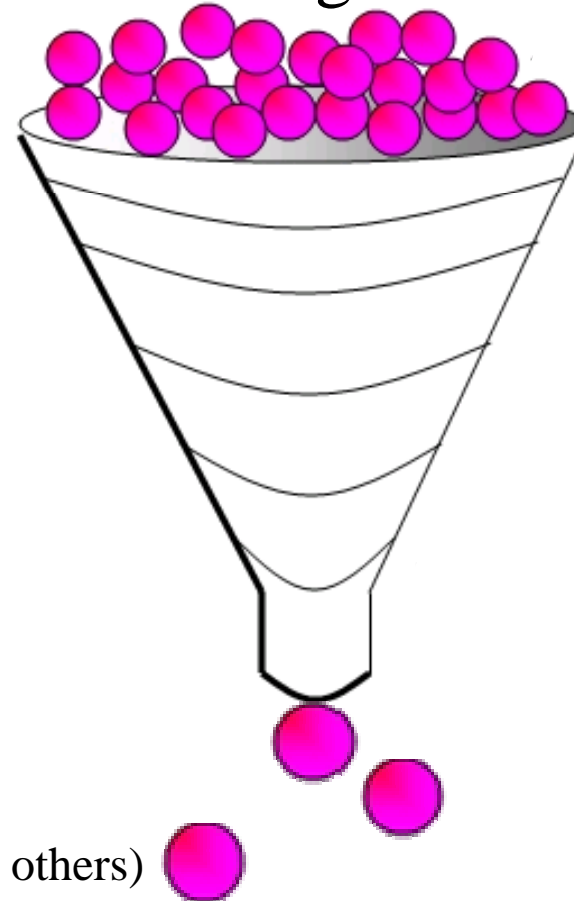
Many Many Tasks: Identifying Potential Drug Targets



Protein
target(s) x

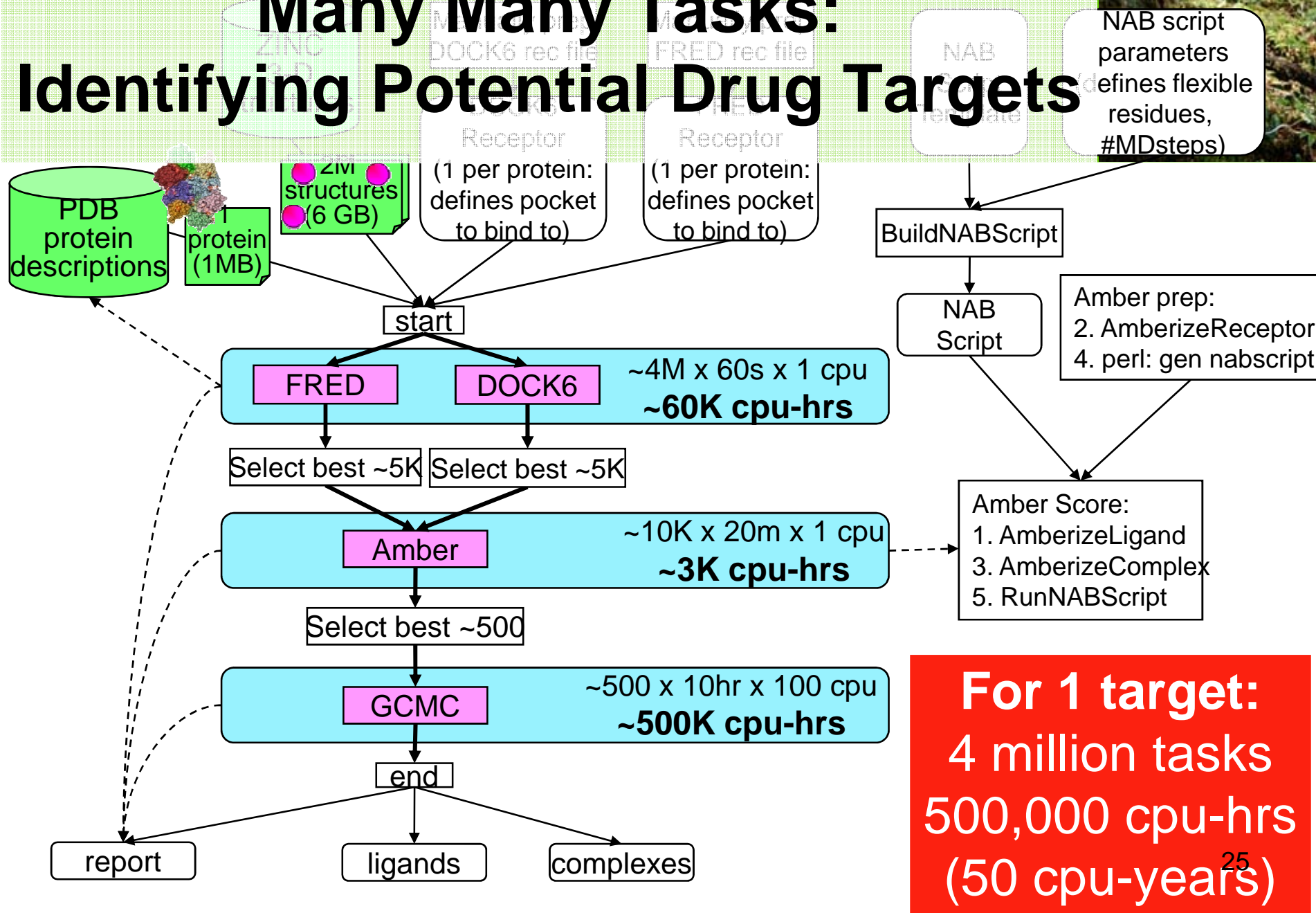


2M+ ligands



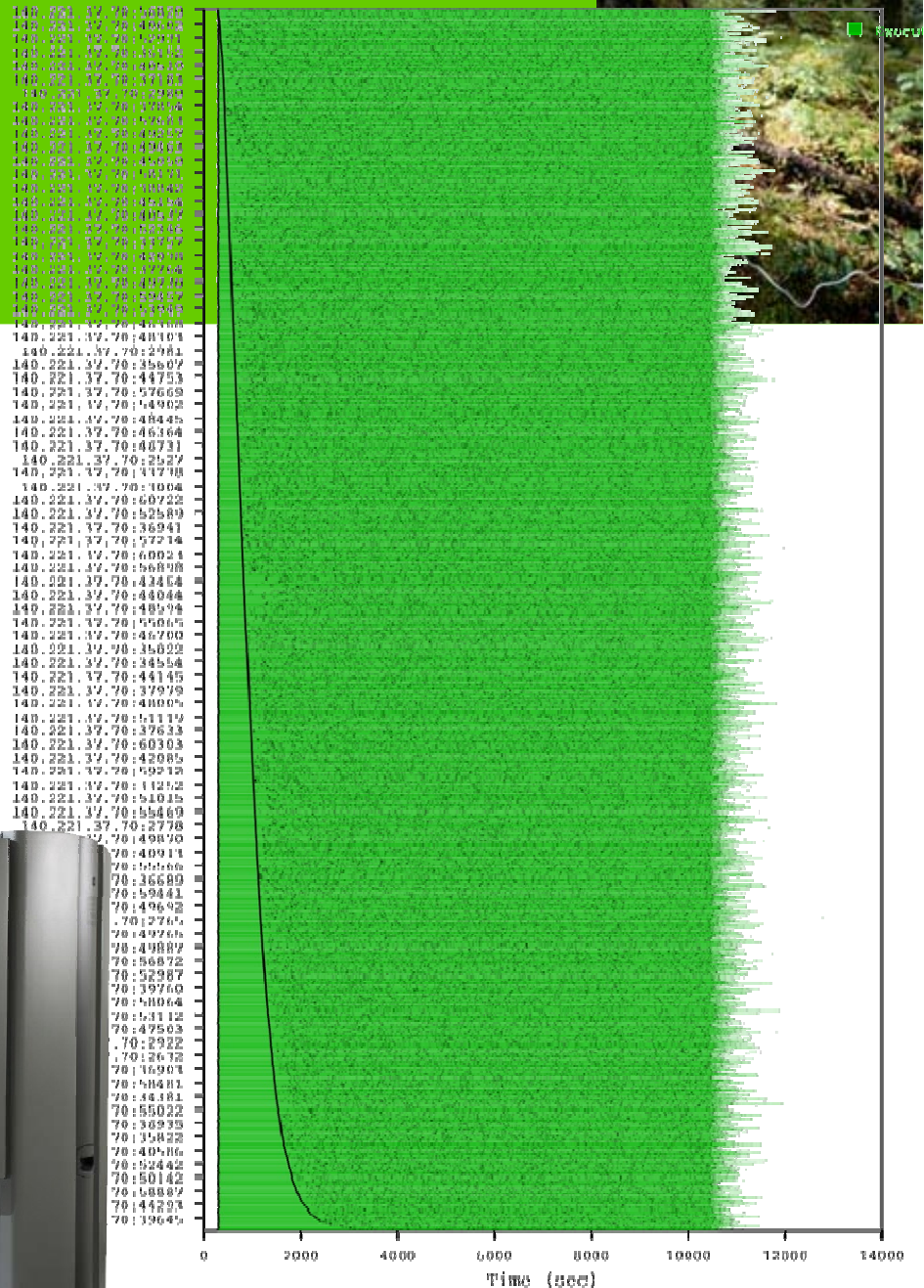
(Mike Kubal, Benoit Roux, and others)

Many Many Tasks: Identifying Potential Drug Targets



DOCK on SiCortex

- CPU cores: 5760
- Tasks: 92160
- Elapsed time: 12821 sec
- Compute time: 1.94 CPU years
- Average task time: 660.3 sec
- Speedup: 5650X (ideal 5760)
- Efficiency: 98.2%



Falkon, a Fast and Efficient Execution Framework for Clusters, Computers

DOCK on the BG/P



CPU cores: 118784

Tasks: 934803

Elapsed time: 2.01 hours

Compute time: 21.43 CPU years

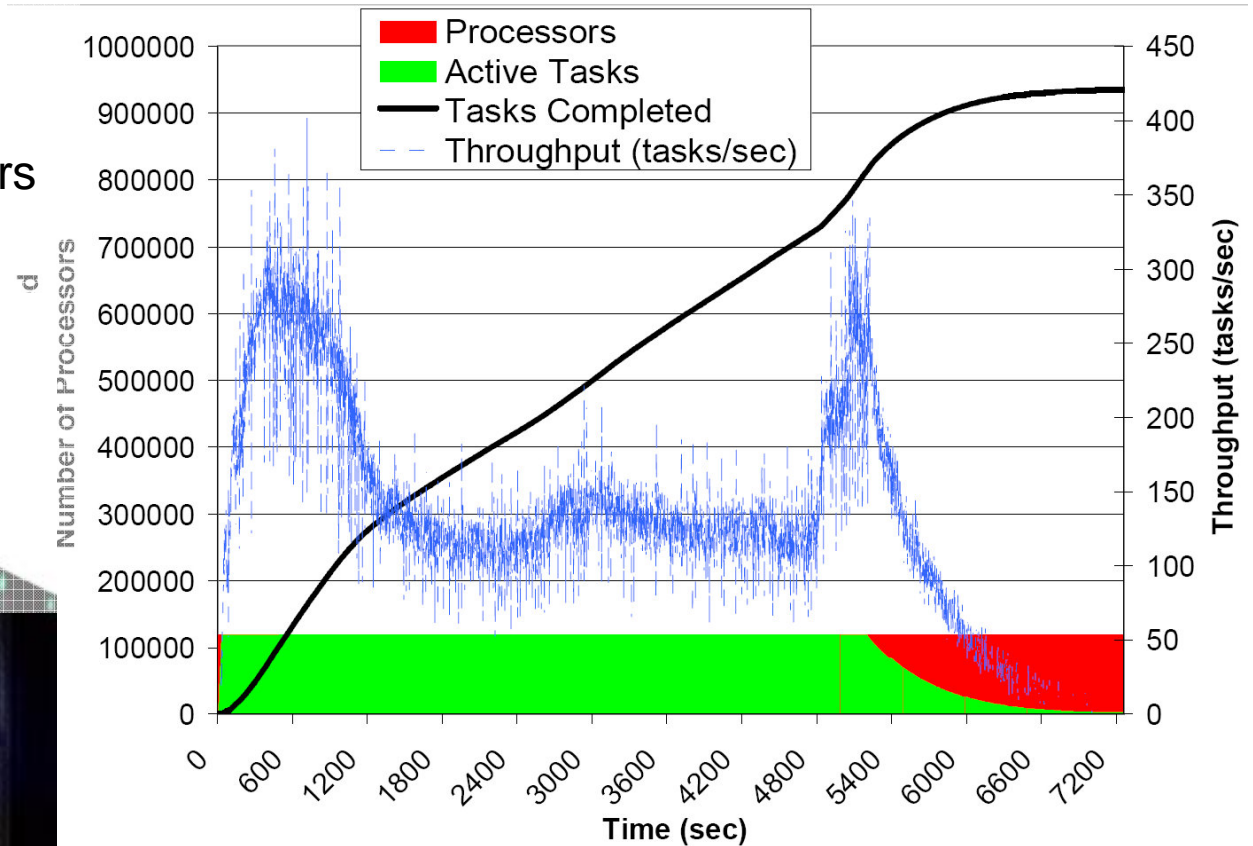
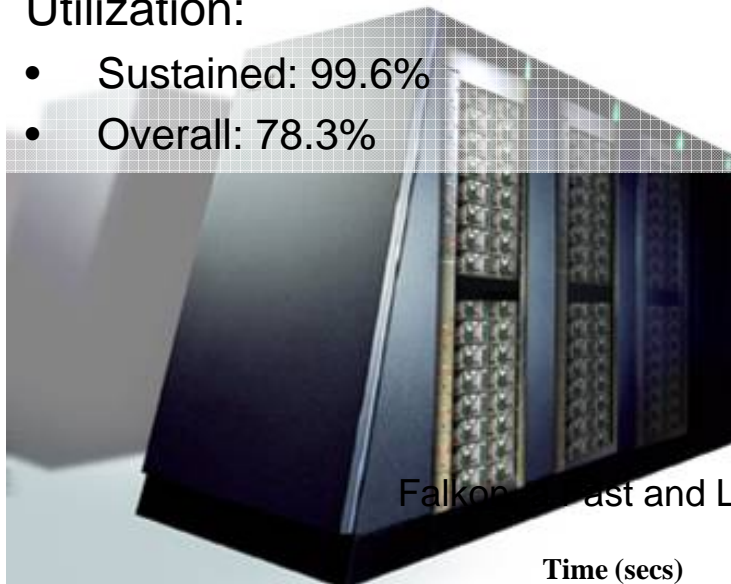
Average task time: 667 sec

Relative Efficiency: 99.7%

(from 16 to 32 racks)

Utilization:

- Sustained: 99.6%
- Overall: 78.3%



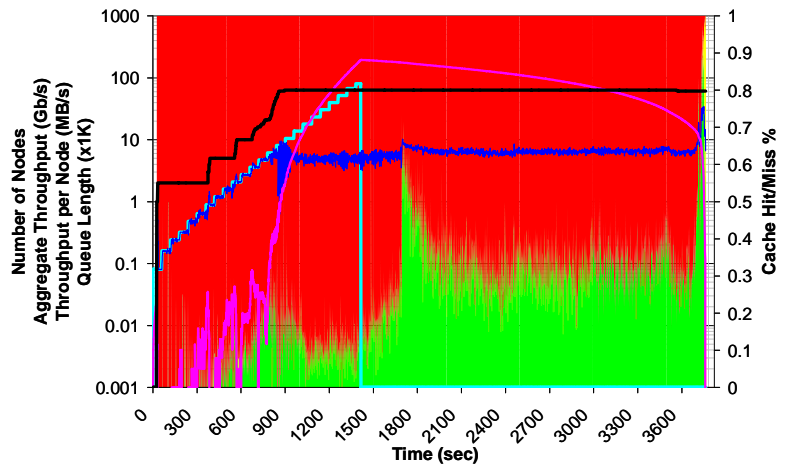
Falcon: Fast and Light-weight task execution framework for Clusters, Grids, and Supercomputers

Data Diffusion



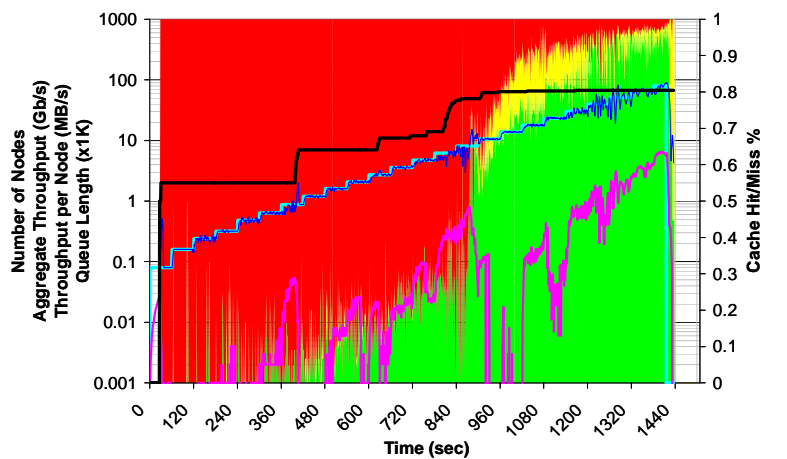
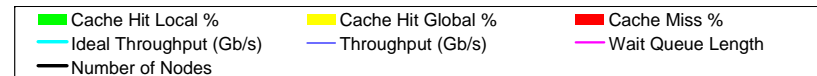
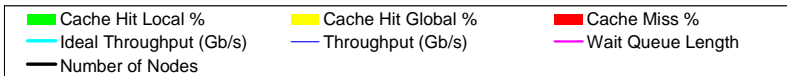
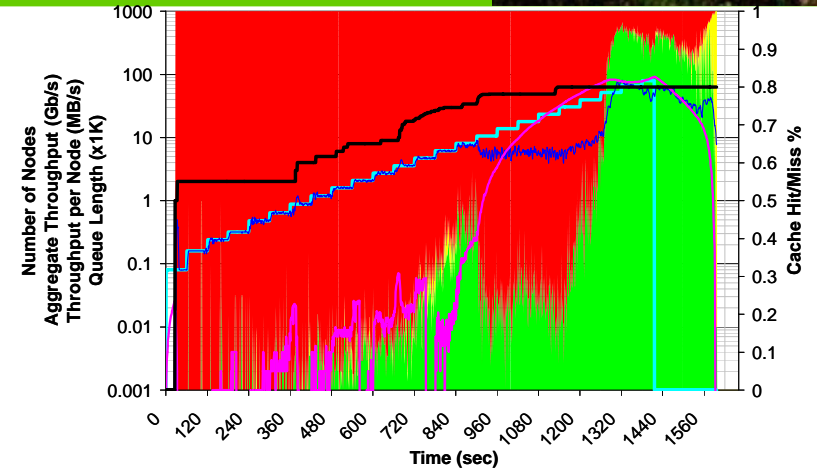
- Considers both data and computations to optimize performance
 - Supports data-aware scheduling
 - Can optimize compute utilization, cache hit performance, or a mixture of the two
- Decrease dependency of a shared file system
 - Theoretical linear scalability with compute resources
 - Significantly increases meta-data creation and/or modification performance
- Central for “data-centric task farm” realization

Data Diffusion: Good-cache-compute



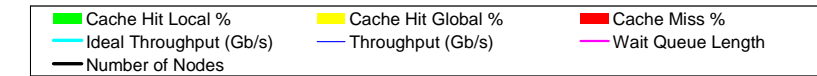
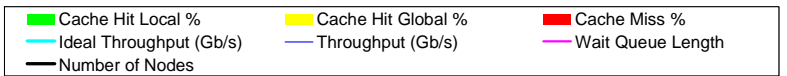
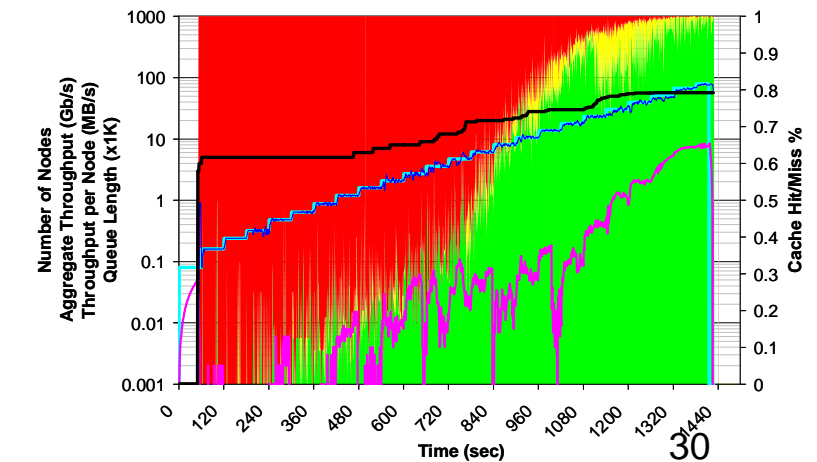
← 1GB

1.5GB →



← 2GB

4GB →



Data Diffusion: Throughput and Response Time

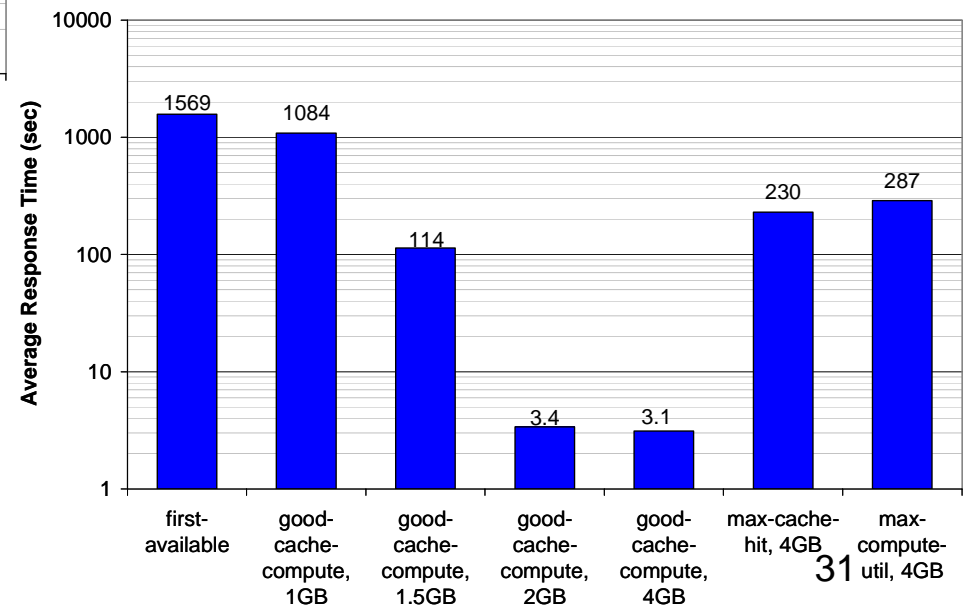


← Throughput:

- Average: 14Gb/s vs 4Gb/s
- Peak: 100Gb/s vs. 6Gb/s

Response Time →

- 3 sec vs 1569 sec → 506X

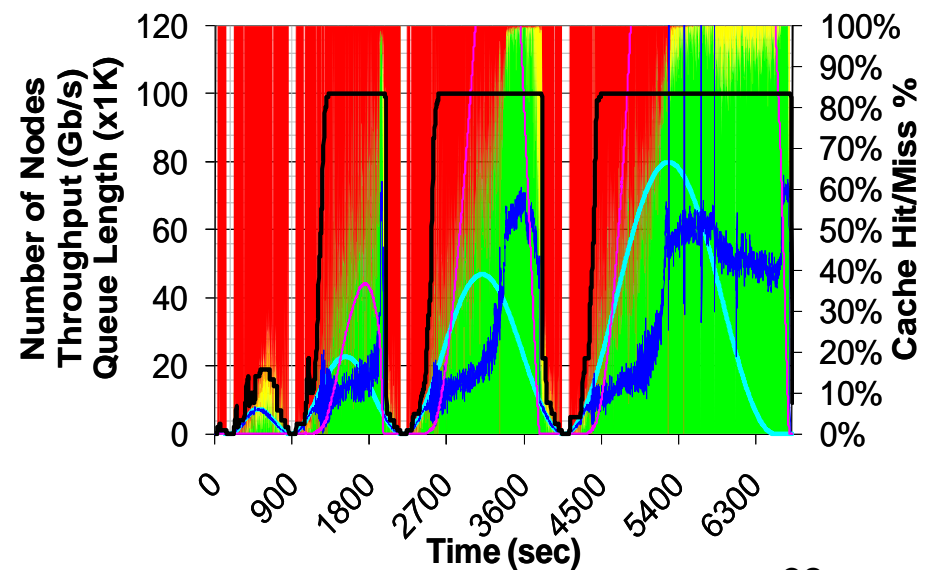
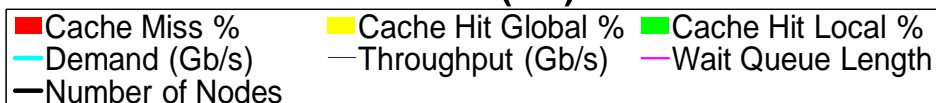
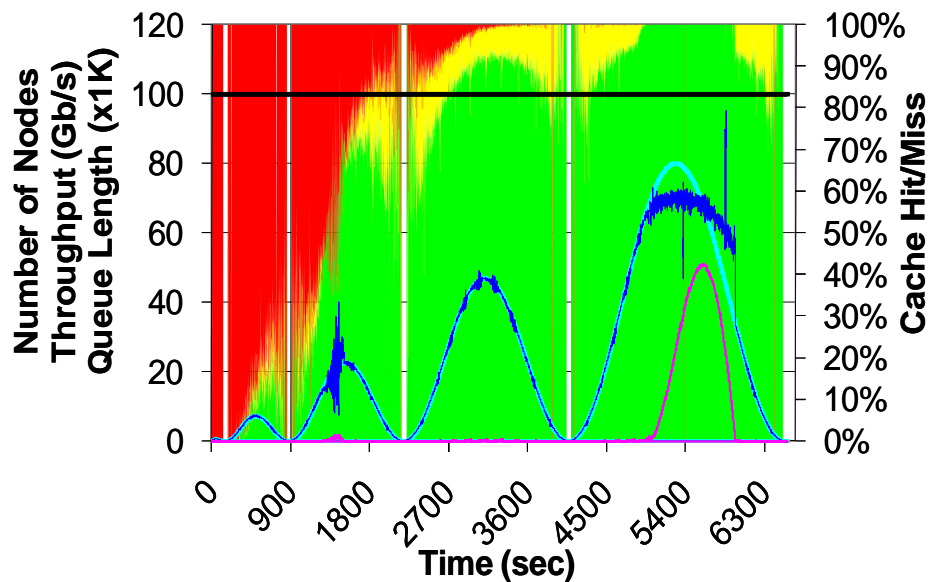


31

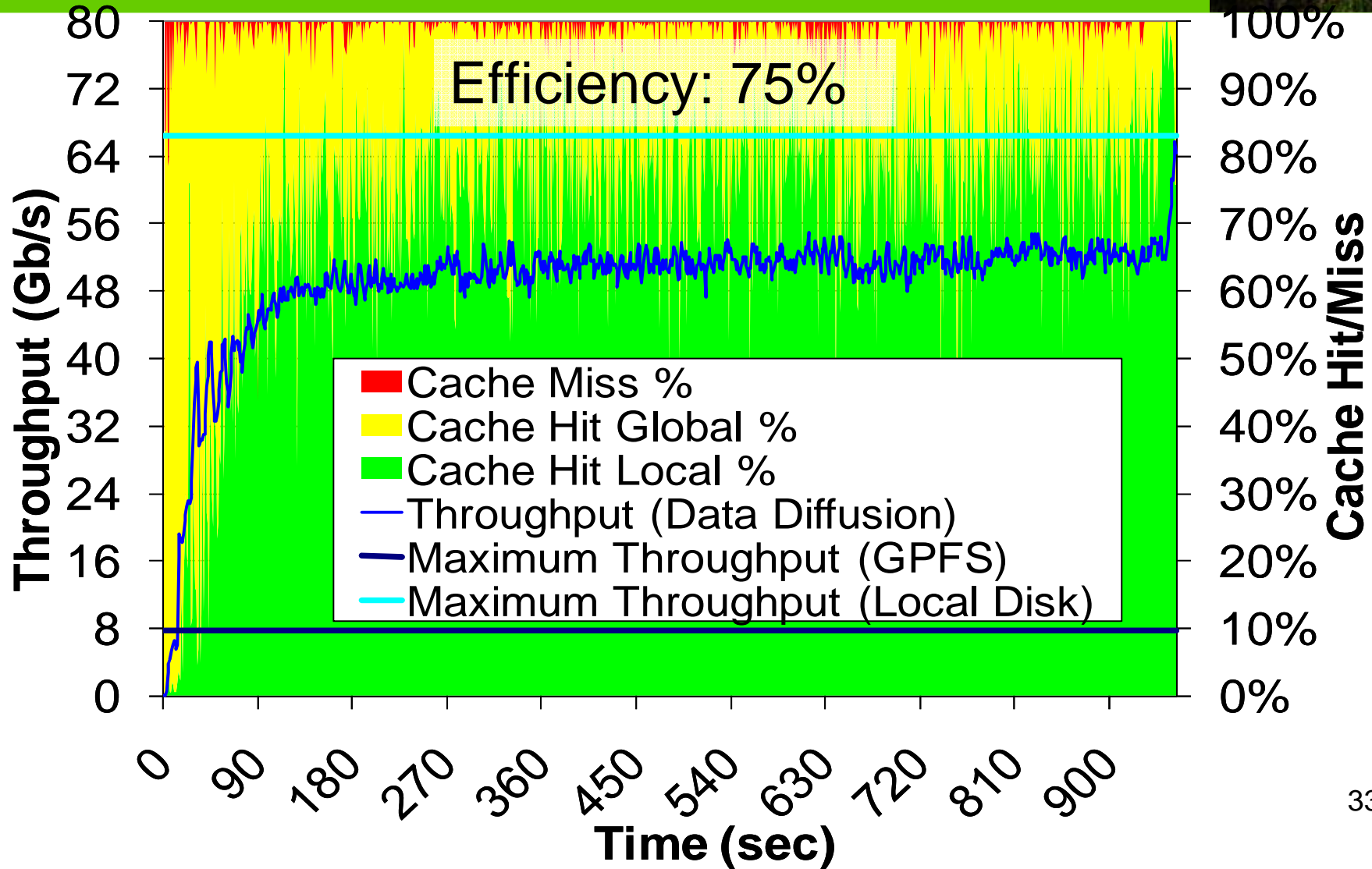
Sin-Wave Workload



- GPFS → 5.7 hrs, ~8Gb/s, 1138 CPU hrs
- DF+SRP → 1.8 hrs, ~25Gb/s, 361 CPU hrs
- DF+DRP → 1.86 hrs, ~24Gb/s, 253 CPU hrs



All-Pairs Workload 500x500 on 200 CPUs



Limitations of Data Diffusion



- Needs Java 1.4+
- Needs IP connectivity between hosts
- Needs local storage (disk, memory, etc)
- Per task workings set must fit in local storage
- Task definition must include input/output files metadata
- Data access patterns: write once, read many

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/>
- Related Projects:
 - Falkon: <http://dev.globus.org/wiki/Incubator/Falkon>
 - Swift: <http://www.ci.uchicago.edu/swift/index.php>
- 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