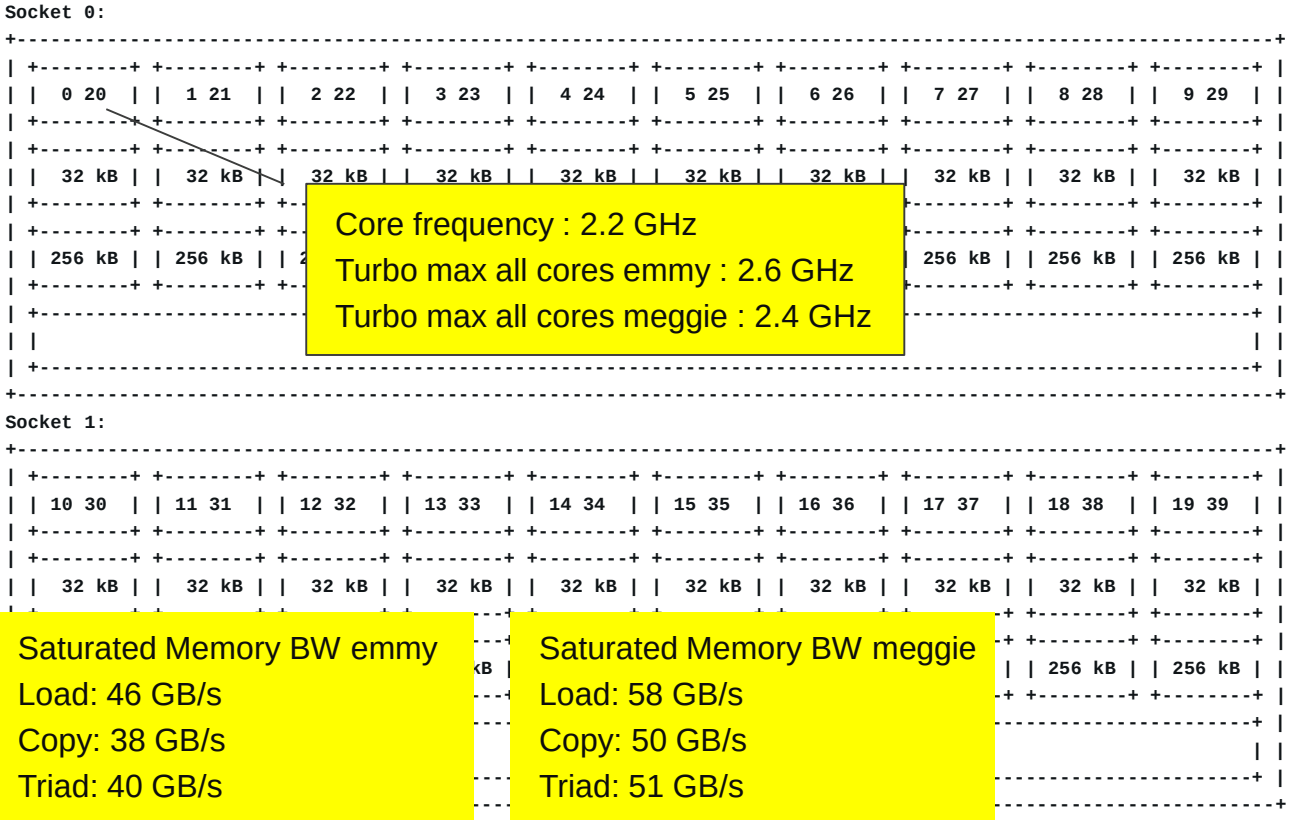
REGIONALES RECHENZENTRUM ERLANGEN [RRZE]

# HPC –Café talk
# How to interpret the Job Performance Monitoring Graphs and what to look for

HPC Services, RRZE

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

# Topology of Erlangen systems: Emmy + Meggie




Core frequency : 2.2 GHz
Turbo max all cores emmy : 2.6 GHz
Turbo max all cores meggie : 2.4 GHz

Saturated Memory BW emmy
Load: 46 GB/s
Copy: 38 GB/s
Triad: 40 GB/s

Saturated Memory BW meggie
Load: 58 GB/s
Copy: 50 GB/s
Triad: 51 GB/s

## Node topology

- 2 sockets
- 10 cores / socket
- 2 HW threads / core

## Memory Hierarchy

- 32kB L1 D-Cache
- 256kB L2 Cache
- 25MB shared L3
- 1 memory domain per socket

# Reminder – Access to performance monitoring

https://www.hpc.rrze.fau.de/HPC-Status/job-info.php

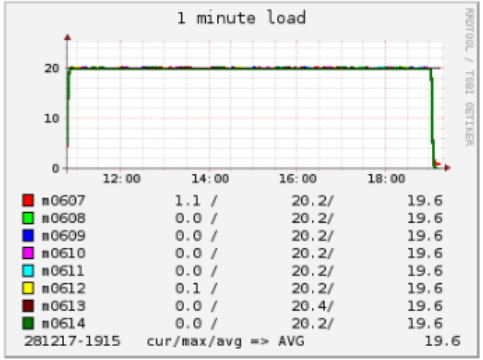
- The required "**AccessKey**" can be found in STDOUT of jobs on Emmy (and Meggie)

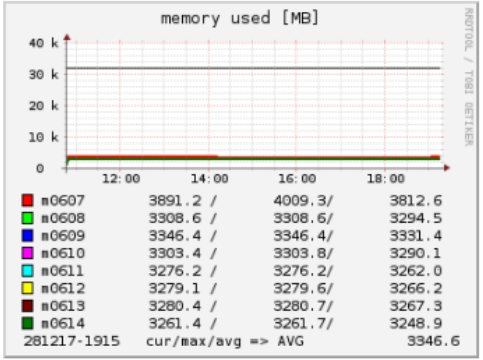
```
=== JOB_STATISTICS ===
=== current date     : Sun Aug 21 15:15:16 CEST 2016
= Job-ID             : 647692.eadm
= Job-Name           : 75
= Queue              : devel
= PBS_O_WORKDIR      : /home/woody/unrz/unrz143/xx-ccm+
= Requested resources: ncpus=1,neednodes=4:ppn=40,nodes=4:ppn=40,walltime=00:30:00
= Used resources     : cput=00:32:18,mem=1515348kb,vmem=15194140kb,walltime=00:03:13
= Node list          : e0108,e0107,e0106,e0105
= AccessKey          : 1e64c726
======================
```

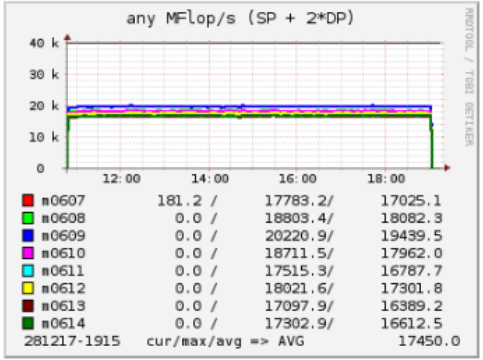
- Monitoring data is typically kept for 2-4 weeks.
  (Currently longer due to preparation of Cluster-2020 proposal)

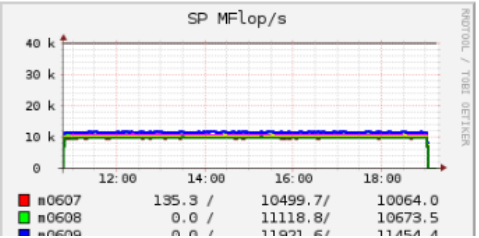
## === Monitoring data ===

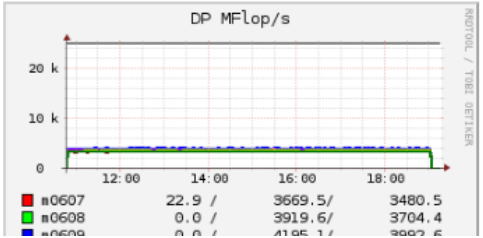
```
elapsed time                                :    30562.0 s
average 1 minute load                       :       19.6
average memory used                         :     3346.6 MB per node
avg any floating point rate (SP + 2*DP)     :    17450.0 MFlop/s per node (SP + 2*DP)
average SP floating point rate              :    10278.0 MFlop/s per node
average DP floating point rate              :     3586.0 MFlop/s per node
average memory bandwidth (read+write)       :    30254.6 MB/s per node
average memory read bandwidth               :    25387.3 MB/s per node
average memory write bandwidth              :     4867.3 MB/s per node
average clock frequency                     :     2366.6 MHz
average chip RAPL-power                     :      112.7 W
average dram RAPL-power                     :       13.3 W
estimated total incoming Ethernet data      :    18961.0 MB aggregated
estimated total outgoing Ethernet data      :     3103.7 MB aggregated
```

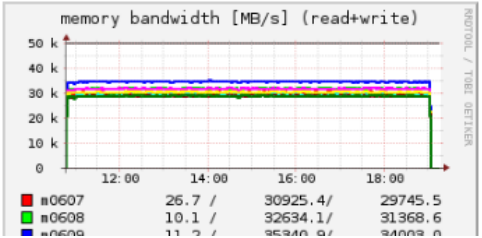



## === job environment ===

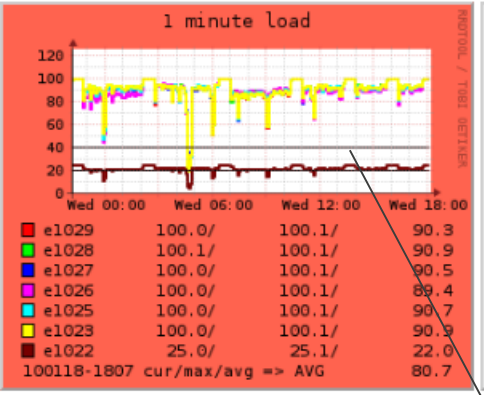
```
SLURM_EXPORT_ENV=NONE
SLURM_JOB_NAME=ccmtest_intel
SLURM_PRIO_PROCESS=0
SLURM_SUBMIT_DIR=/lxfs/unrz/unrz143/STAR-TE
SLURM_SUBMIT_HOST=meggie1.rrze.uni-erlangen
SLURM_UMASK=0077
SLURM_GET_USER_ENV=1
```

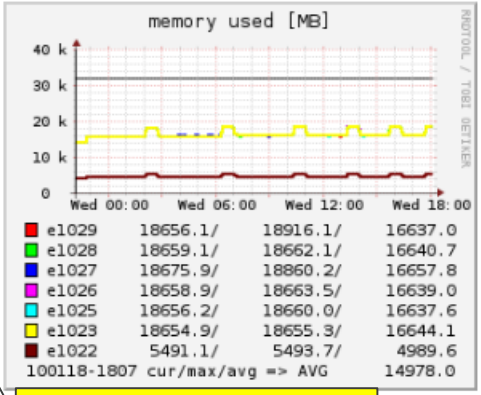
## === job script ===

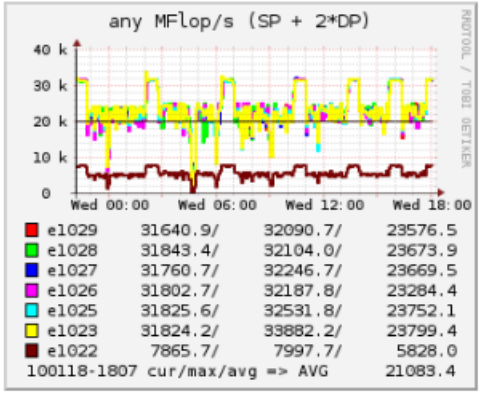
```
#!/bin/bash -l
#SBATCH --job-name=ccmtest_intel
#SBATCH --partition=work
#SBATCH --nodes=8
#SBATCH --time=9:00:00
#SBATCH --export=NONE

# star-ccm+ arguments
CCMARGS="test.java -load test.sim"

# specify the time you want to have to save
```

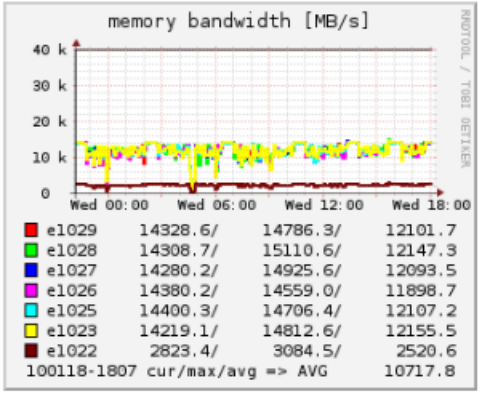
# Example: Resource Oversubscription

- Load of ~100 on some nodes??



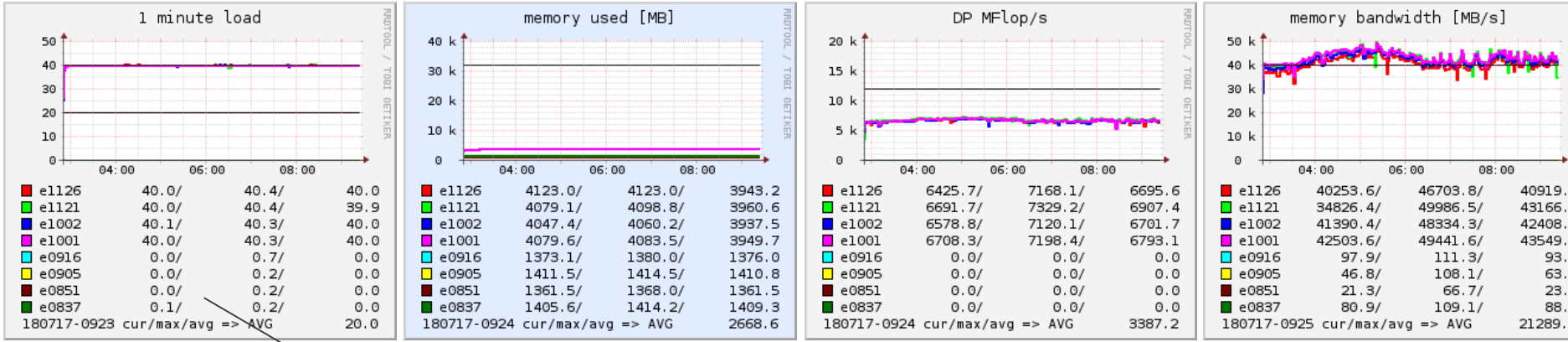




Load should usually not exceed black line

# Example: Only processes on half of the nodes

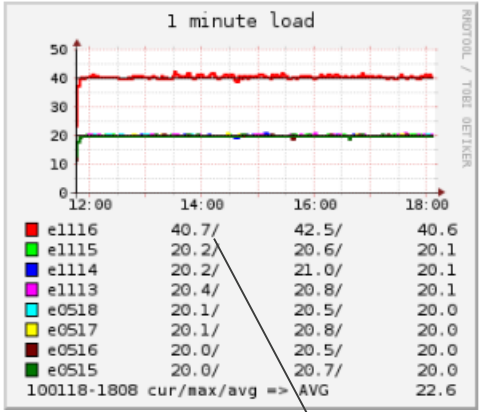
- Wrong resource allocation for mpirun

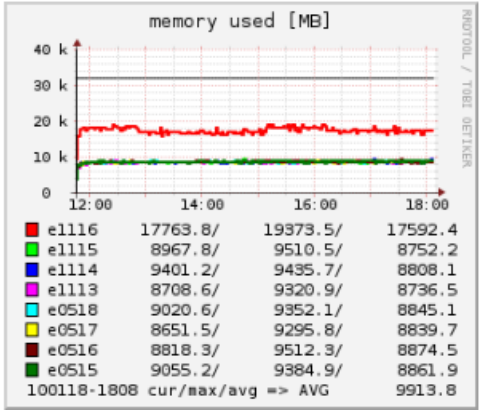


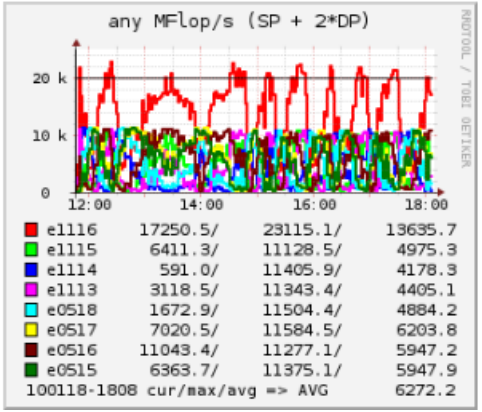

Specify processes per node if no SMT should be used

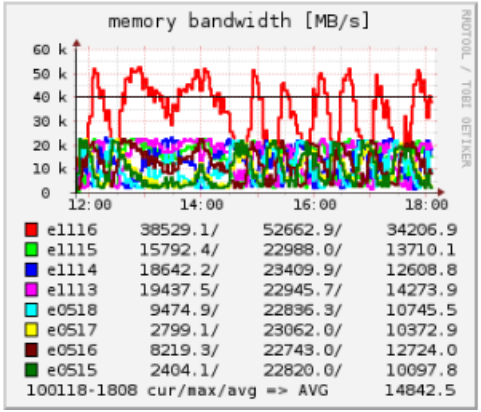
# Example: Using SMT on subset of nodes

- 1st node overloaded



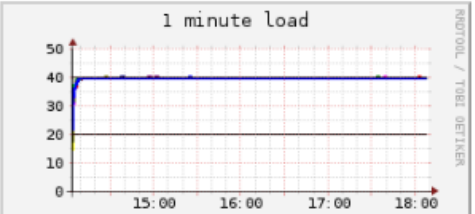




Only one node uses SMT

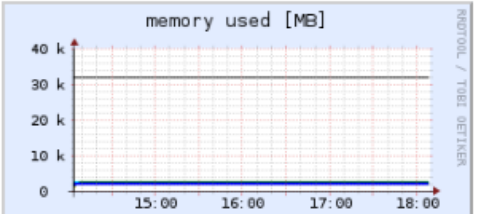
# Example: Load Imbalance

- Severe work imbalance

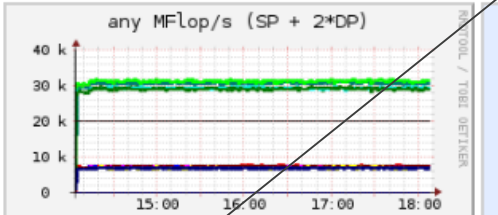
Multiple node performance classes



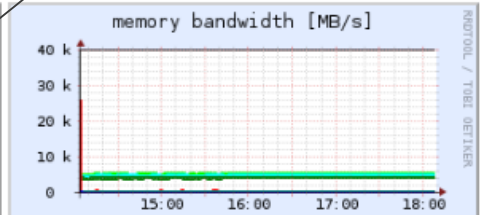

| | | | |
|---|---|---|---|
| e1021 | 40.0/ | 40.3/ | 39.8 |
| e1020 | 40.0/ | 40.2/ | 39.8 |
| e1019 | 40.0/ | 40.2/ | 39.8 |
| e1018 | 40.0/ | 40.2/ | 39.8 |
| e1017 | 40.0/ | 40.2/ | 39.8 |
| e1016 | 40.1/ | 40.2/ | 39.8 |
| e0806 | 40.1/ | 40.2/ | 39.8 |
| e0805 | 40.0/ | 40.2/ | 39.8 |
| e0804 | 40.0/ | 40.5/ | 39.8 |
| e0803 | 40.0/ | 40.3/ | 39.8 |
| e0802 | 40.0/ | 40.2/ | 39.7 |
| e0656 | 40.0/ | 40.4/ | 39.7 |
| e0655 | 40.0/ | 40.2/ | 39.7 |
| e0654 | 40.1/ | 40.2/ | 39.7 |
| e0653 | 40.0/ | 40.3/ | 39.7 |
| e0652 | 40.1/ | 40.3/ | 39.7 |
| e0651 | 40.1/ | 40.2/ | 39.7 |




| | | | |
|---|---|---|---|
| e1021 | 2689.9/ | 2695.5/ | 2655.8 |
| e1020 | 2527.0/ | 2535.1/ | 2520.7 |
| e1019 | 2538.5/ | 2542.8/ | 2528.2 |
| e1018 | 2550.2/ | 2558.3/ | 2541.3 |
| e1017 | 2580.2/ | 2587.1/ | 2566.9 |
| e1016 | 2526.5/ | 2530.7/ | 2516.7 |
| e0806 | 2635.9/ | 2643.2/ | 2624.6 |
| e0805 | 2734.0/ | 2742.7/ | 2715.3 |
| e0804 | 2633.4/ | 2639.1/ | 2619.5 |
| e0803 | 2628.4/ | 2635.7/ | 2616.6 |
| e0802 | 2734.4/ | 2742.1/ | 2720.8 |
| e0656 | 2597.6/ | 2604.0/ | 2581.7 |
| e0655 | 2637.5/ | 2637.5/ | 2614.3 |
| e0654 | 2741.1/ | 2749.8/ | 2726.5 |
| e0653 | 2605.8/ | 2607.5/ | 2591.9 |
| e0652 | 2583.5/ | 2587.9/ | 2570.1 |
| e0651 | 2699.1/ | 2702.0/ | 2680.2 |




| | | | |
|---|---|---|---|
| e1021 | 47.3/ | 494.6/ | 31.2 |
| e1020 | 47.1/ | 712.7/ | 32.1 |
| e1019 | 48.3/ | 846.0/ | 32.9 |
| e1018 | 49.1/ | 744.4/ | 35.6 |
| e1017 | 52.0/ | 735.4/ | 45.8 |
| e1016 | 48.3/ | 672.5/ | 34.6 |
| e0806 | 7326.9/ | 7467.5/ | 7216.3 |
| e0805 | 29114.7/ | 29901.5/ | 29194.6 |
| e0804 | 6917.9/ | 7119.2/ | 6958.6 |
| e0803 | 7285.8/ | 7539.7/ | 7312.4 |
| e0802 | 30632.4/ | 31662.3/ | 30722.8 |
| e0656 | 7141.1/ | 7249.3/ | 7032.1 |
| e0655 | 7765.0/ | 7844.6/ | 7610.4 |
| e0654 | 31665.5/ | 32101.4/ | 31160.1 |
| e0653 | 7572.0/ | 7619.8/ | 7393.4 |
| e0652 | 7444.1/ | 7596.4/ | 7362.7 |
| e0651 | 29747.7/ | 30004.8/ | 29248.9 |

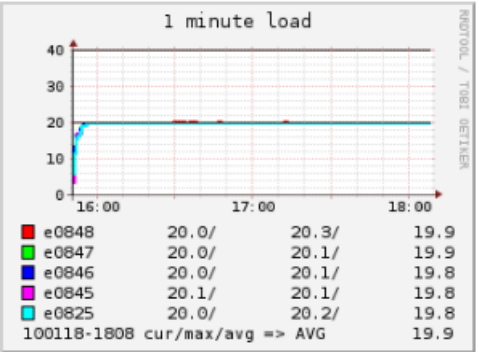


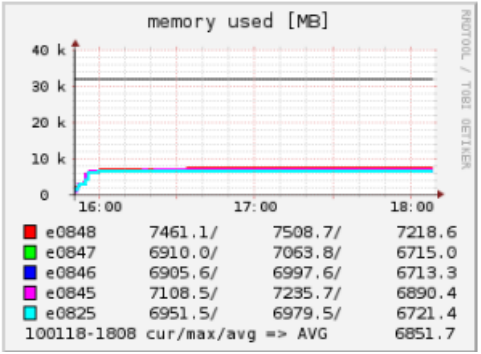

| | | | |
|---|---|---|---|
| e1021 | 155.2/ | 14434.0/ | 213.0 |
| e1020 | 152.8/ | 29639.0/ | 279.0 |
| e1019 | 157.9/ | 10750.2/ | 200.8 |
| e1018 | 181.5/ | 17423.8/ | 253.8 |
| e1017 | 209.7/ | 23046.6/ | 307.6 |
| e1016 | 156.2/ | 20113.9/ | 238.0 |
| e0806 | 401.1/ | 424.4/ | 402.7 |
| e0805 | 4326.8/ | 4385.3/ | 4228.5 |
| e0804 | 316.5/ | 326.2/ | 312.9 |
| e0803 | 569.1/ | 574.5/ | 558.2 |
| e0802 | 5014.1/ | 5085.6/ | 4897.4 |
| e0656 | 459.4/ | 477.8/ | 458.8 |
| e0655 | 588.5/ | 604.0/ | 586.1 |
| e0654 | 5563.7/ | 5612.4/ | 5410.5 |
| e0653 | 452.9/ | 476.2/ | 456.5 |
| e0652 | 530.3/ | 550.2/ | 525.2 |
| e0651 | 5063.0/ | 5198.5/ | 4999.1 |

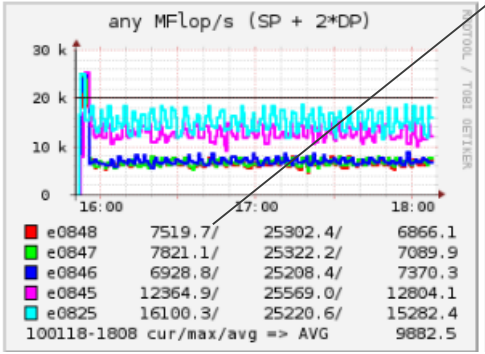
# Examples

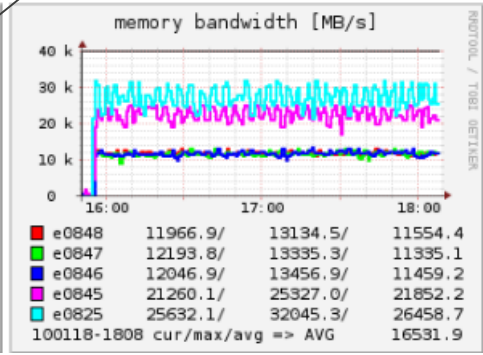
- Severe work imbalance

Multiple node performance classes



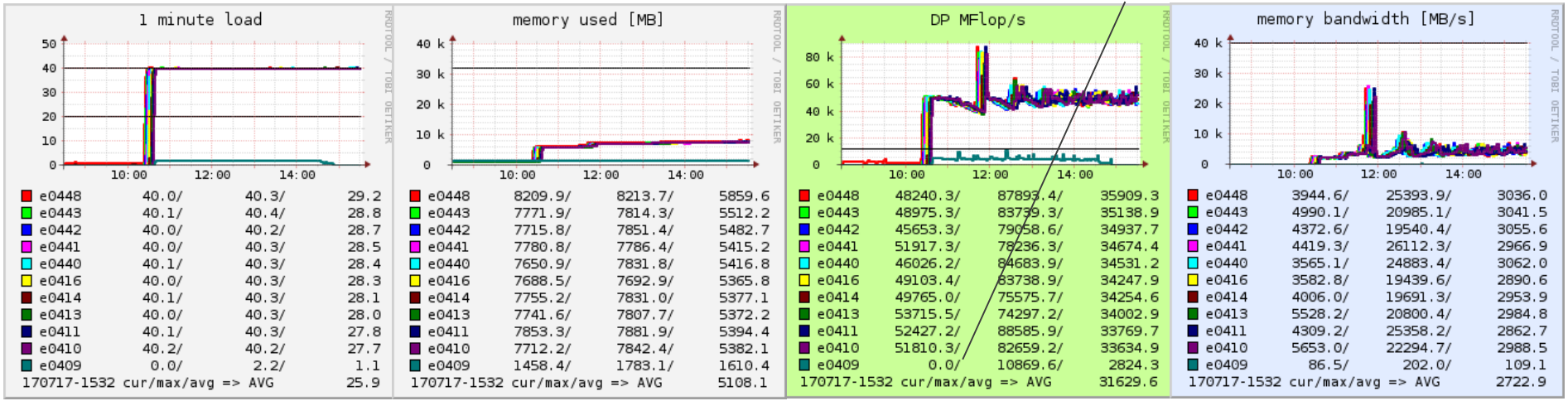



# Examples

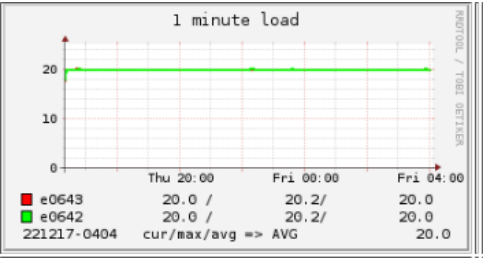
- Severe work imbalance

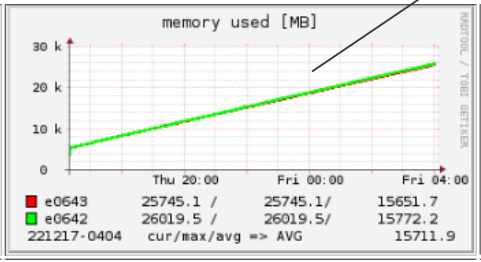
On last node only
→ OK if you have to use specific process number!

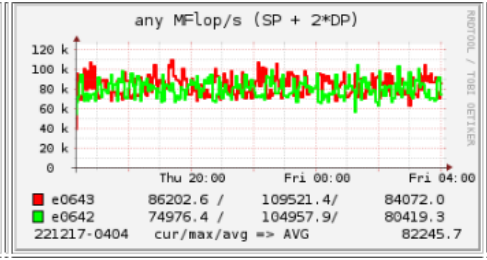


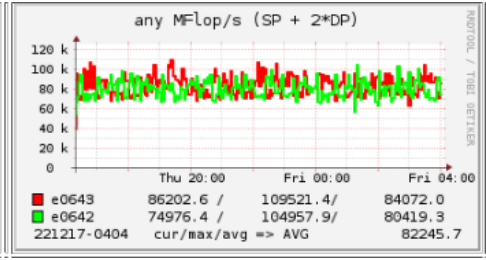
# Examples: Memory leak

Memory used increases with runtime



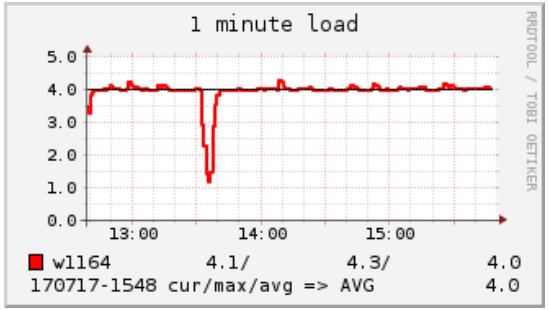



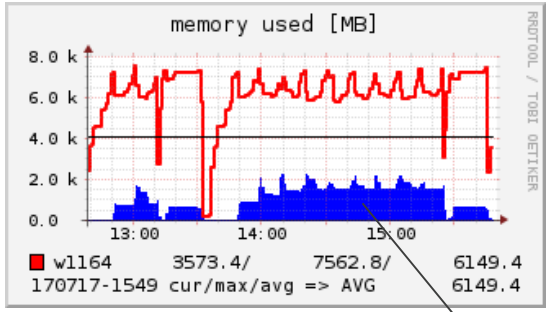



# Examples: Swapping node

- Memory of node exceeded




Swap in use

# Wrap up: Things to look for

- Load does not fit # cores or # SMT threads
- Low resource utilization (flops_any, mem_bw)
- Idle nodes
- Multiple performance classes
- Monotonically increasing memory capacity / swapping
- Long sequential Pre-/Postprocessing Phases

# Questions!?

HPC@RRZE

**https://hpc.fau.de**