PRACE T7.2-5IP

Towards Energy-efficient Exascale Computing: A Use-case Applying READEX to Alya

PRACE @ ISC 2018

Ramón Martínez Carreras
Irish Centre for High-End Computing (ICHEC)

ramon.carreras@ichec.ie

Agenda

1. PRACE-5IP T7.2
2. READEX
3. EoCoE
4. Results
5. Summary

1 PRACE-5IP T7.2

1.1. Context

1.1. Context

- Looking ahead to future Exascale systems.
- Investigating new **tools**/**techniques** by applying them to important **applications**.
- Task 7.2 implemented through mini-projects between CoEs and European Exascale projects (primarily FET-HPC).
- Period: Jan 2017 – Mar 2019.

2 READEX

European Commission | Horizon 2020 European Union funding for Research & Innovation

READEX Runtime Exploitation of Application Dynamism for Energy-efficient eXascale computing

FET-HPC project, launched 09/2015

www.readex.eu

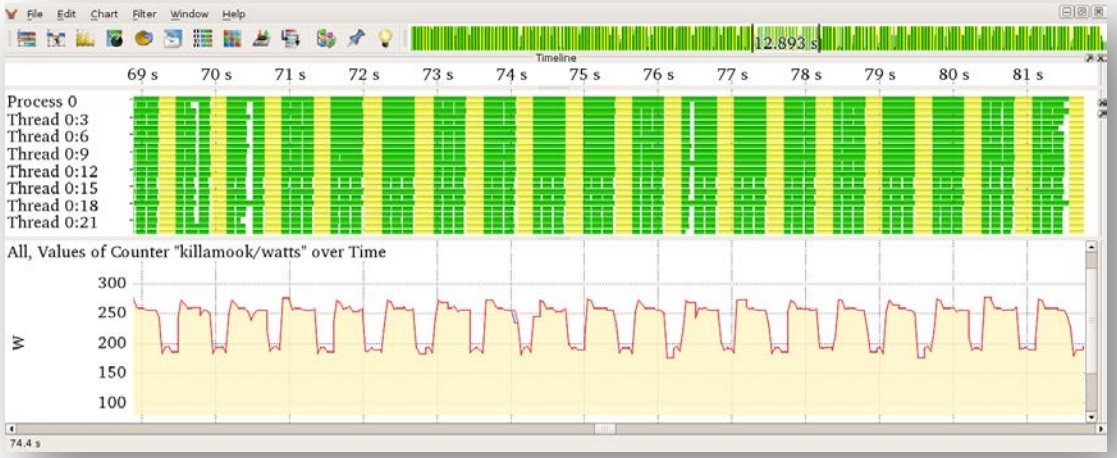
2.1. Consortium

- Technische Universität Dresden/ZIH (Coordinator)
- Norges Teknisk-Naturvitenskapelige Universitet
- Technische Universität München
- IT4Innovations, VSB-Technical University of Ostrava
- Irish Centre for High-End Computing
- Intel Corporation SAS
- Gesellschaft für numerische Simulation mbH

2.2. Objectives

HPC applications exhibit **dynamic behaviour** at runtime:

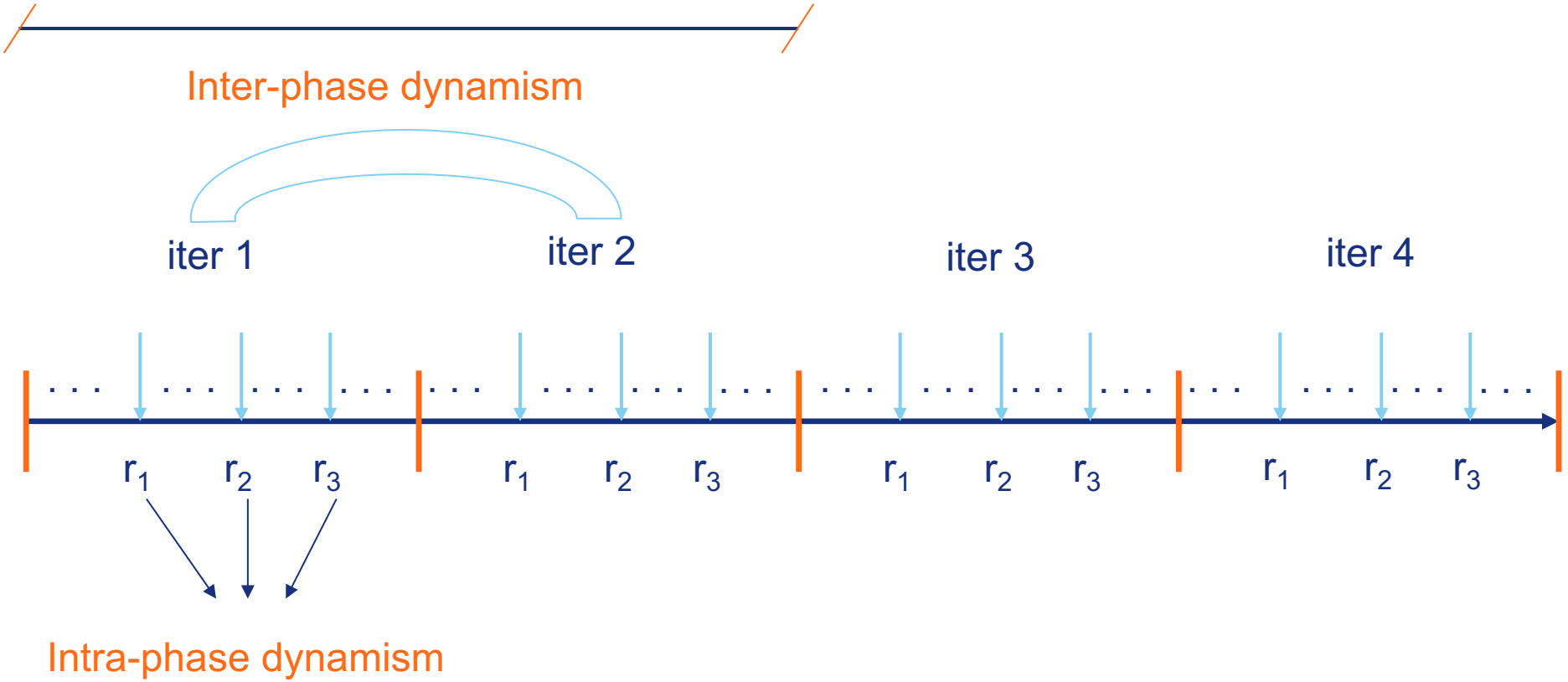
- Dynamic workload characteristics
- Dynamic resource requirements



READEX

Runtime Exploitation of Application Dynamism for Energy-efficient eXascale computing

2.3. Application Dynamism

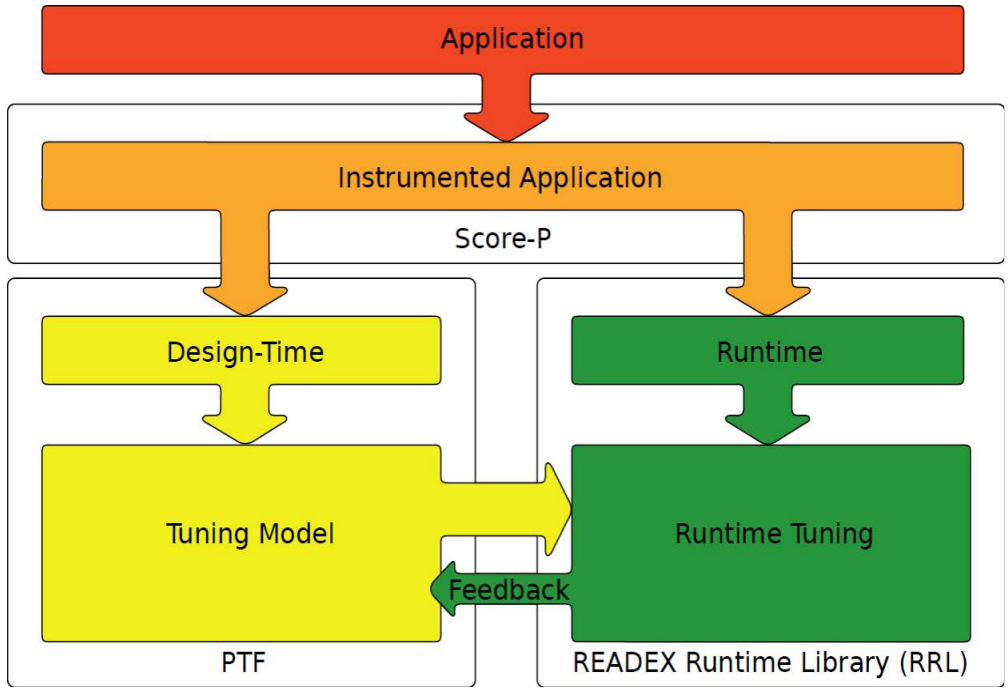


2.4. Tuning Parameters

Level	Tuning aspect	Tuning Parameter	Scope
Hardware Parameters	CPU Frequency Controls	Core frequency (DVFS) Uncore frequency Energy Performance Bias (EPB)	Core Socket Socket
System Software Parameters	OpenMP Parallelism	Dynamic Concurrency Throttling Workload scheduling algorithm	Process Process
Application Parameters	User-specified code-paths	… Decomposition Compiler type & compiler parameter Type of iterative & direct solvers Preprocessing of stiffness & coarse problem matrices …	 Application Application Application Application

2.5. READEX Workflow

1. **Application instrumentation**
   - Use Score-P to instrument (identify) key application regions
2. **Dynamism detection**
   - Check the application for dynamism to benefit from READEX tuning
3. **Design Time Analysis (DTA)**
   - Perform experimental runs
   - Identify optimal configurations and potential energy savings
   - Create *tuning model*
4. **Runtime Application Tuning (RAT)**
   - Production run of application
   - Use tuning model to apply optimal configurations
   - Update tuning model using *calibration* for unseen scenarios

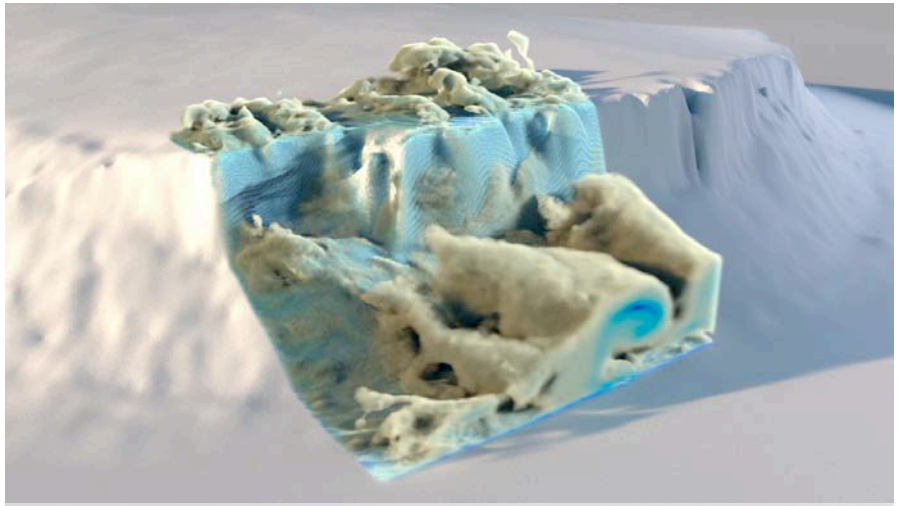


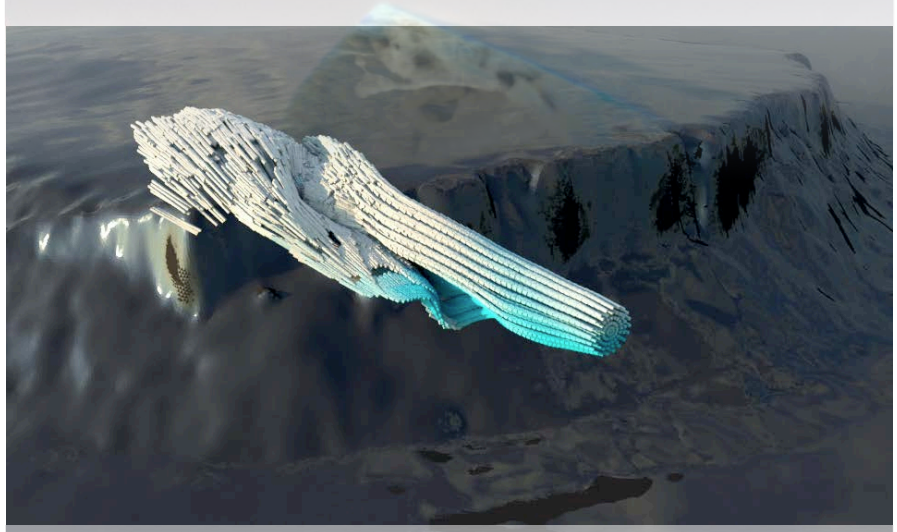
3 EoCoE

3.1. BSC

E-CoE : **Oriented Centre of Excellence**

- Alya for wind resource assessment. Code validation for complex terrain with forest.
- Alya performs node level optimisation. For eg., 30% CPU time reduction for matrix assembly.
- Implementation and testing of external direct (Mumps and Pastix) and iterative solvers (AGMG, Maphys and PSBlas).

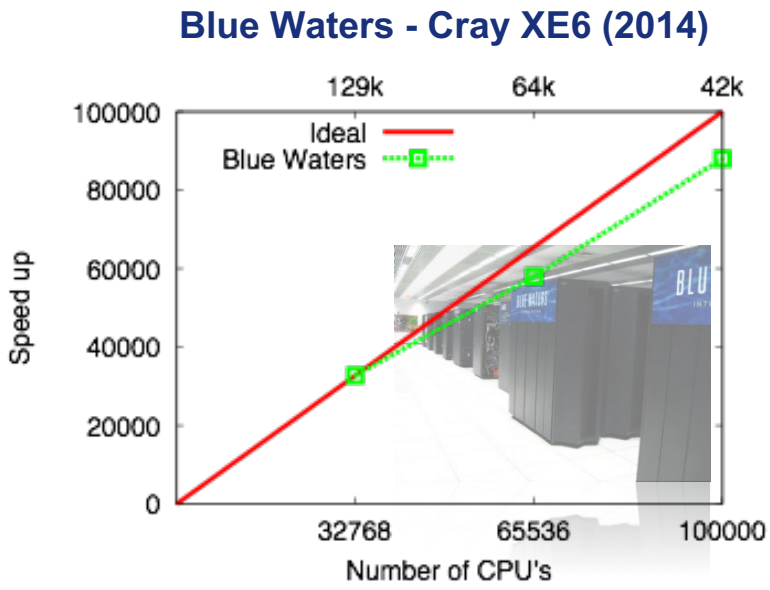




3.2. Alya: High Performance Computational Mechanics

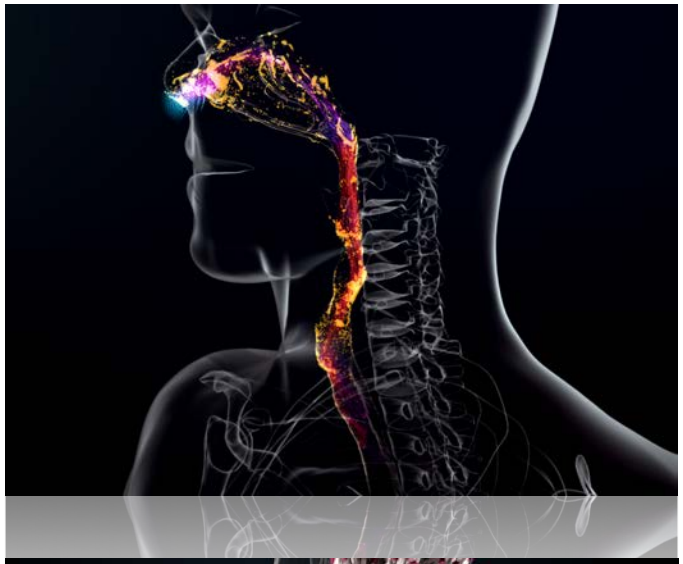
- Finite element, Spectral element, Virtual element
- Low dissipation schemes, turbulence modelling
- Multiphysics, loose and strong couplings
- Multi-code coupling
- Non-conforming meshes
- Fluid, solid, heat transfer, radiation, chemistry, particle transport, electro-magnetism

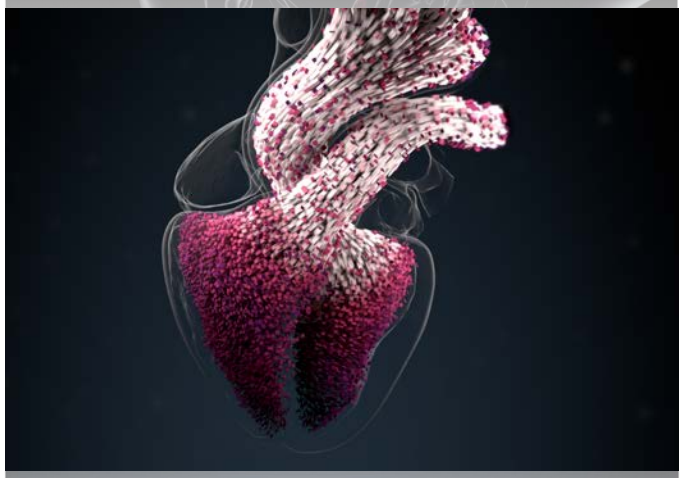
- Hybrid MPI/OpenMP
- MPI: parallel and adaptive partitioning
- OpenMP: loop and task parallelisms
- Dynamic load balance
- Ported to GPU (Cuda/OpenACC)
- Co-execution CPU/GPU



3.3. Alya: Industry Driven

- Present in the Unified European Application Benchmark Suite (out of 12 codes)
- Present in the PRACE Accelerator Benchmark Suite
- Applications in:
  - Biomechanics: cardiovascular and respiratory system
  - Renewable energies
  - Smart cities: air quality in cities
  - Aeronautics: full aircraft and turbines
  - Vehicles





3.4. Alya: Application Parameters for Tuning

Application parameter	Description
VECTOR_SIZE	Size of chunks of elements assembled at the same time (used for vectorization and GPU).
GROUPS	Number of groups used in the Deflated Conjugate Gradient to solve SPD equations (e.g. pressure equation).
KRYLOV	Size of the Krylov subspace used in GMRES solver (e.g. momentum equations, solid mechanics).
RENUMBERING	Node renumbering used on Sparse Matrix Vector Products (SpMV).

4 Results

4.1. Experiments

4.2. Current Results

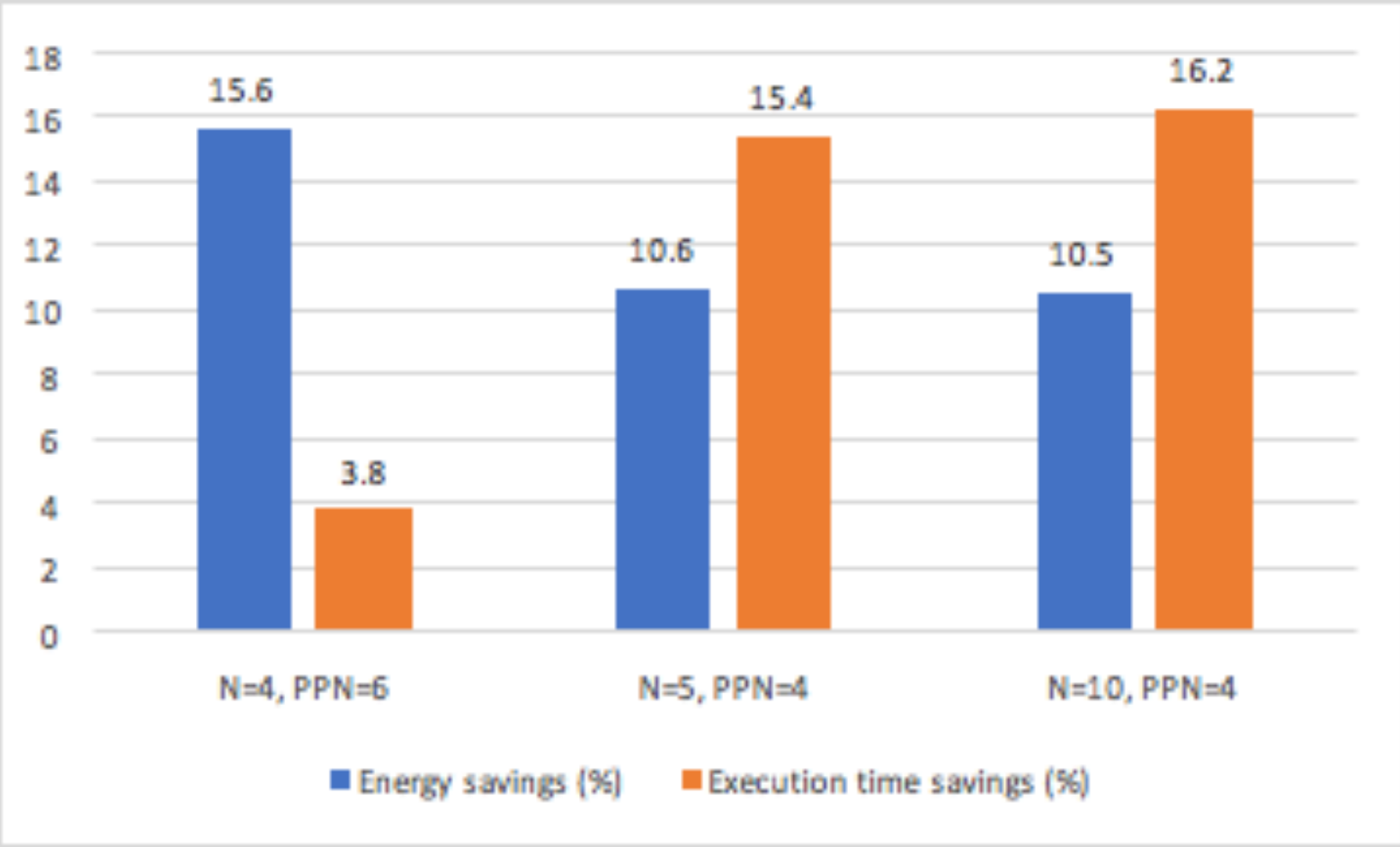
4.1. Experiments

- Taurus cluster @TU Dresden
  - Intel Xeon E5-2680 V3 (dual-socket 24-core Haswell)
- Experiments:
  - Nodes = 4-20
  - MPI processes = 5-80
  - Representative application input (run time about 6.825 min)

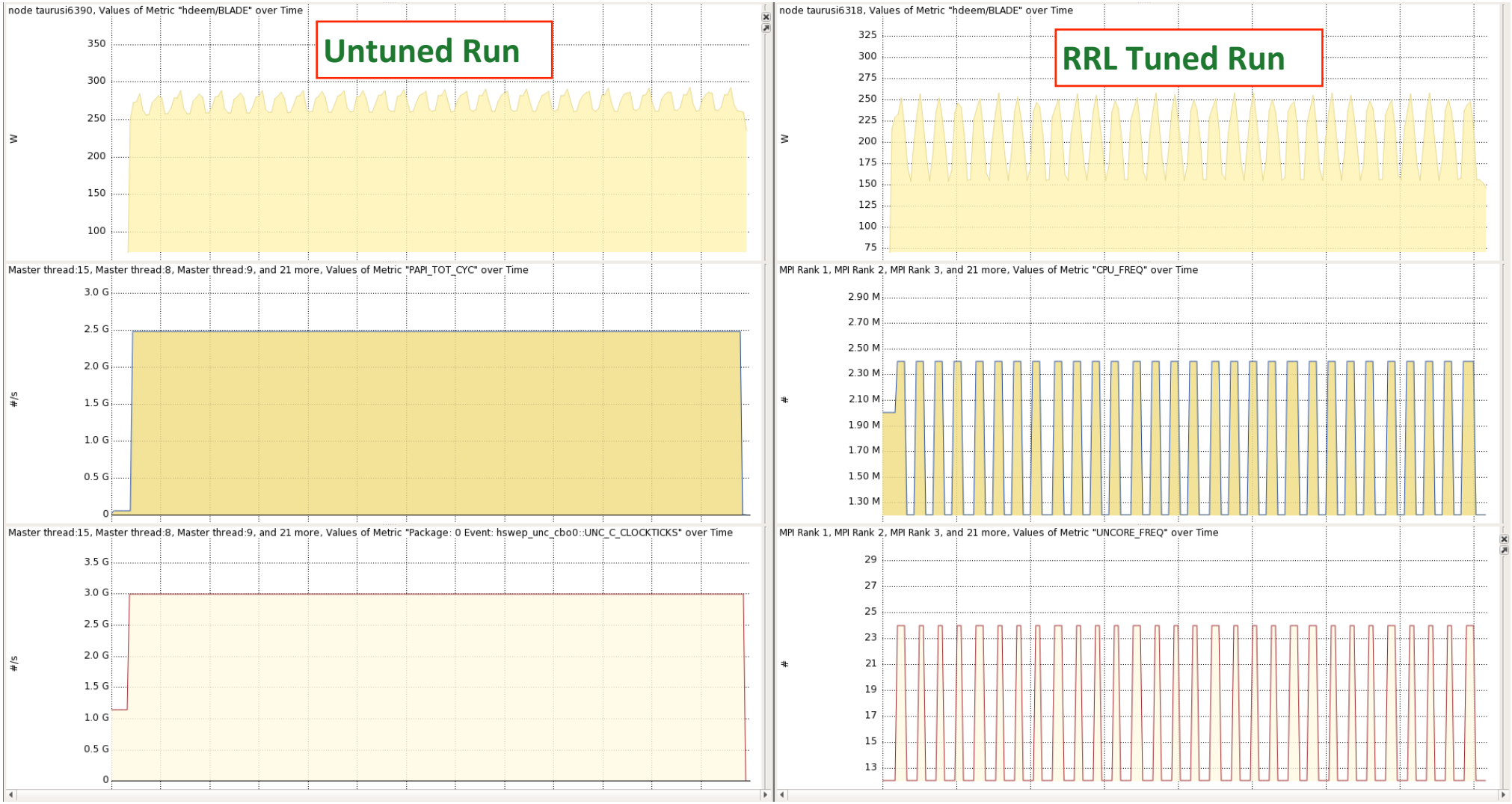
4.2. Results

```
1.  Significant regions are:
2.      DEFLCG
3.      MATRICES_GRADIENT_DIVERGENCE
4.      MATRICES_LAPLACIAN
5.      NSI_DOMMAS
6.      NSI_EIGEN_TIME_STEP_ELEMENT_OPERATIONS
7.      PAR_BARRIER
8.      PAR_MIN_4_s

10. Significant region information
11. ==============================
12. Region name                 Min(t)          Max(t)       Time           Time Dev.(%Reg) Ops/L3miss      Weight(%Phase)
13. NSI_EIGEN_TIME_STEP_ELEME   0.555             0.572        5.571          0.0            863              3
14. MATRICES_GRADIENT_DIVERGE   0.003             0.003        0.003          0.0              0              0
15. MATRICES_LAPLACIAN          0.003             0.003        0.003          0.0              0              0
16. PAR_MIN_4_s                 6.794             6.794        6.794          0.0           1912              4
17. PAR_BARRIER                 0.000             4.787      138.102          0.0           3776             82
18. DEFLCG                      0.291             0.452       10.132          0.0              7              6
19. NSI_DOMMAS                  0.201             0.202        2.015          0.2            523              1

21. Phase information
22. =================
23. Min                 Max                 Mean                Time                Dev.(% Phase)       Dyn.(% Phase)
24. 16.1013             22.0763             16.7625             167.625             0                   35.6448

26. threshold time variation (percent of mean region time): 10.000000
27. threshold compute intensity deviation (#ops/L3 miss): 10.000000
28. threshold region importance (percent of phase exec. time): 10.000000

30. SUMMARY:
31. ========

33. Inter-phase dynamism due to variation of the execution time of phases

35. No intra-phase dynamism due to time variation

37. No intra-phase dynamism due to compute intensity variation
```

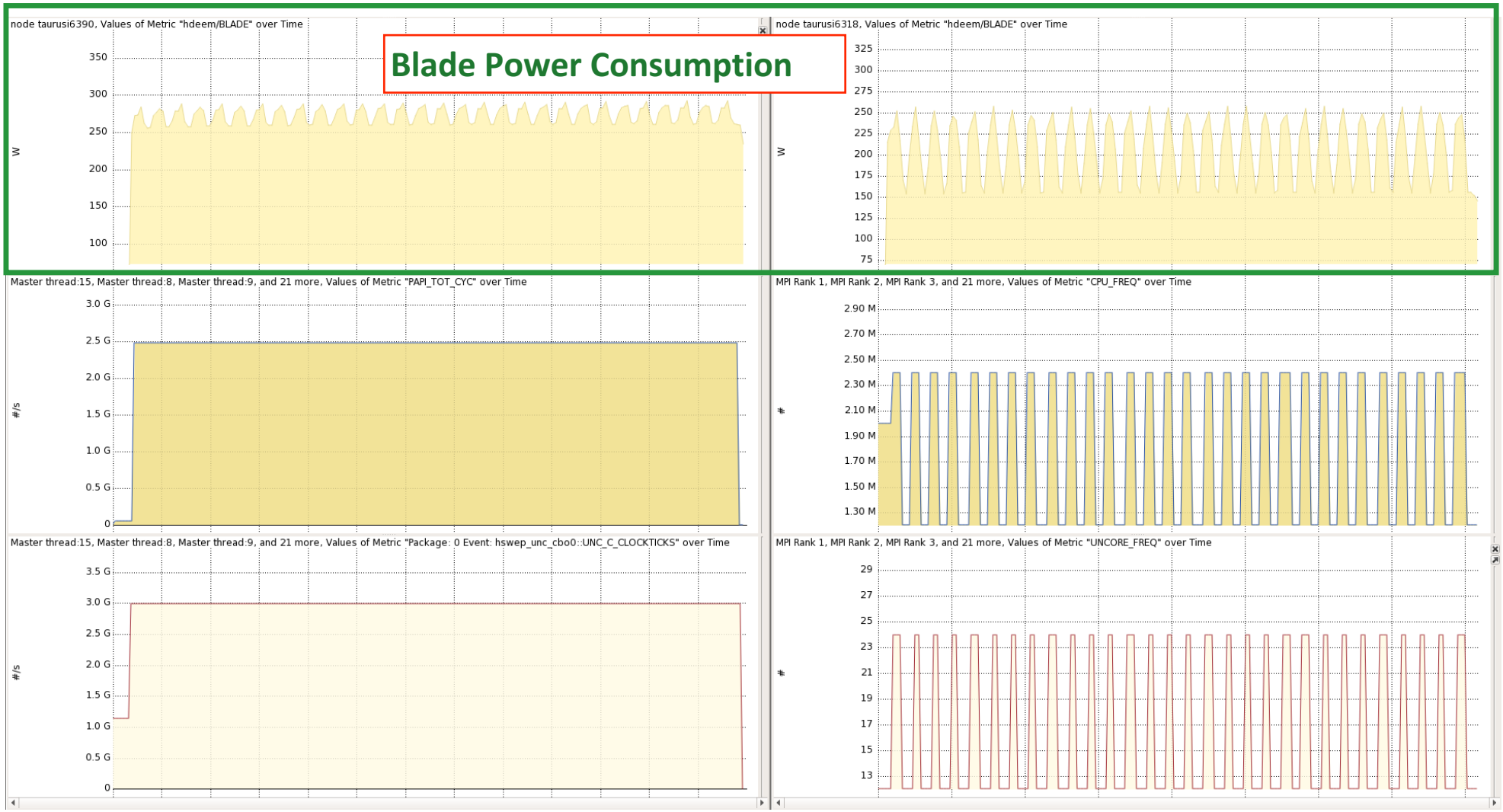
4.2. Results


18
16
14
12
10
8
6
4
2
0
15.6
3.8
10.6
15.4
10.5
16.2
N=4, PPN=6
N=5, PPN=4
N=10, PPN=4
Energy savings (%)
Execution time savings (%)

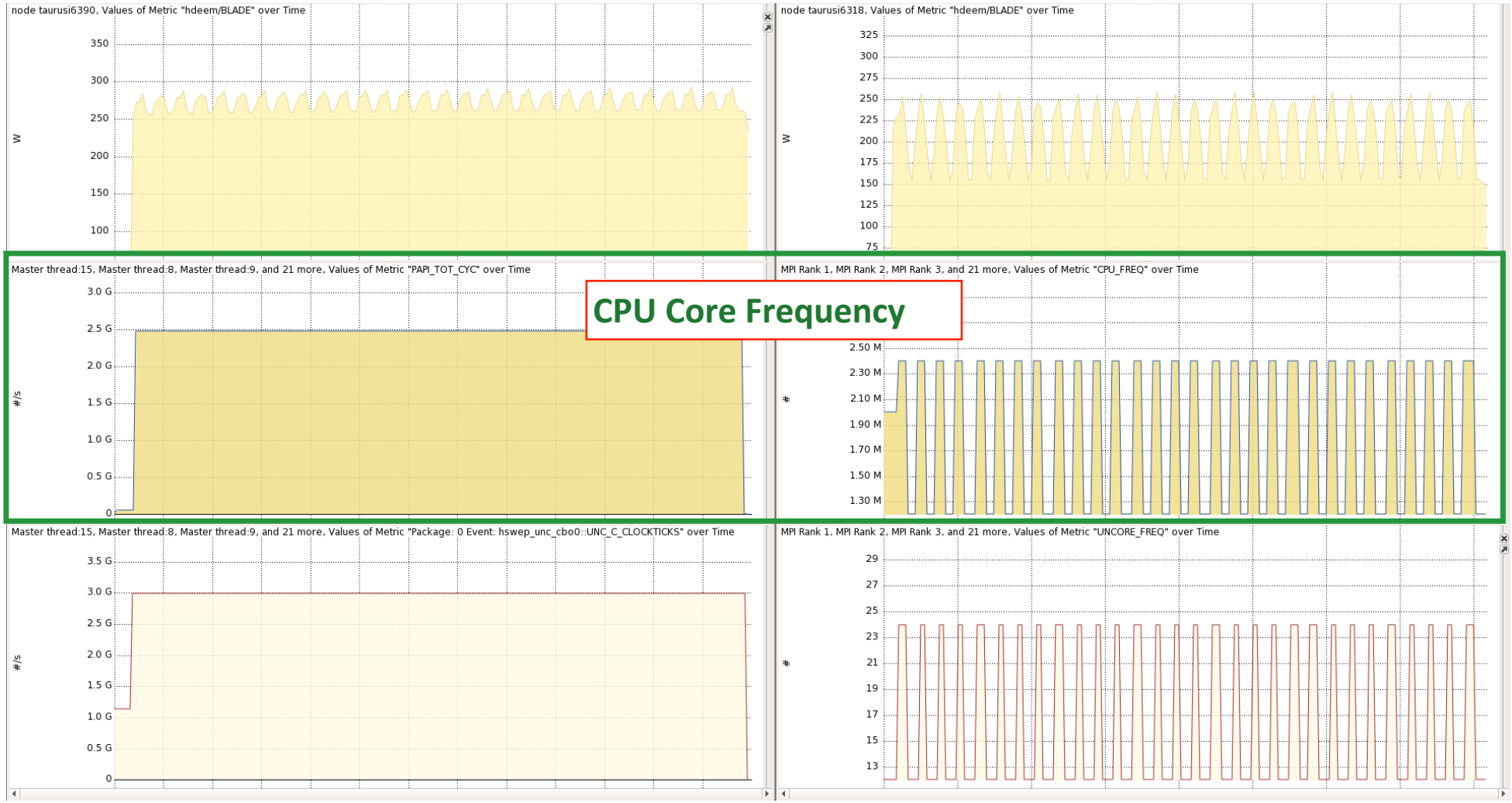
4.2. Results


node taurusi6390, Values of Metric "hdeem/BLADE" over Time
Untuned Run
W
350
300
250
200
150
100
node taurusi6318, Values of Metric "hdeem/BLADE" over Time
RRL Tuned Run
W
325
300
275
250
225
200
175
150
125
100
75
Master thread:15, Master thread:8, Master thread:9, and 21 more, Values of Metric "PAPI_TOT_CYC" over Time
#/s
3.0 G
2.5 G
2.0 G
1.5 G
1.0 G
0.5 G
0
MPI Rank 1, MPI Rank 2, MPI Rank 3, and 21 more, Values of Metric "CPU_FREQ" over Time
#
2.90 M
2.70 M
2.50 M
2.30 M
2.10 M
1.90 M
1.70 M
1.50 M
1.30 M
Master thread:15, Master thread:8, Master thread:9, and 21 more, Values of Metric "Package: 0 Event: hswep_unc_cbo0::UNC_C_CLOCKTICKS" over Time
#/s
3.5 G
3.0 G
2.5 G
2.0 G
1.5 G
1.0 G
0.5 G
0
MPI Rank 1, MPI Rank 2, MPI Rank 3, and 21 more, Values of Metric "UNCORE_FREQ" over Time
#
29
27
25
23
21
19
17
15
13

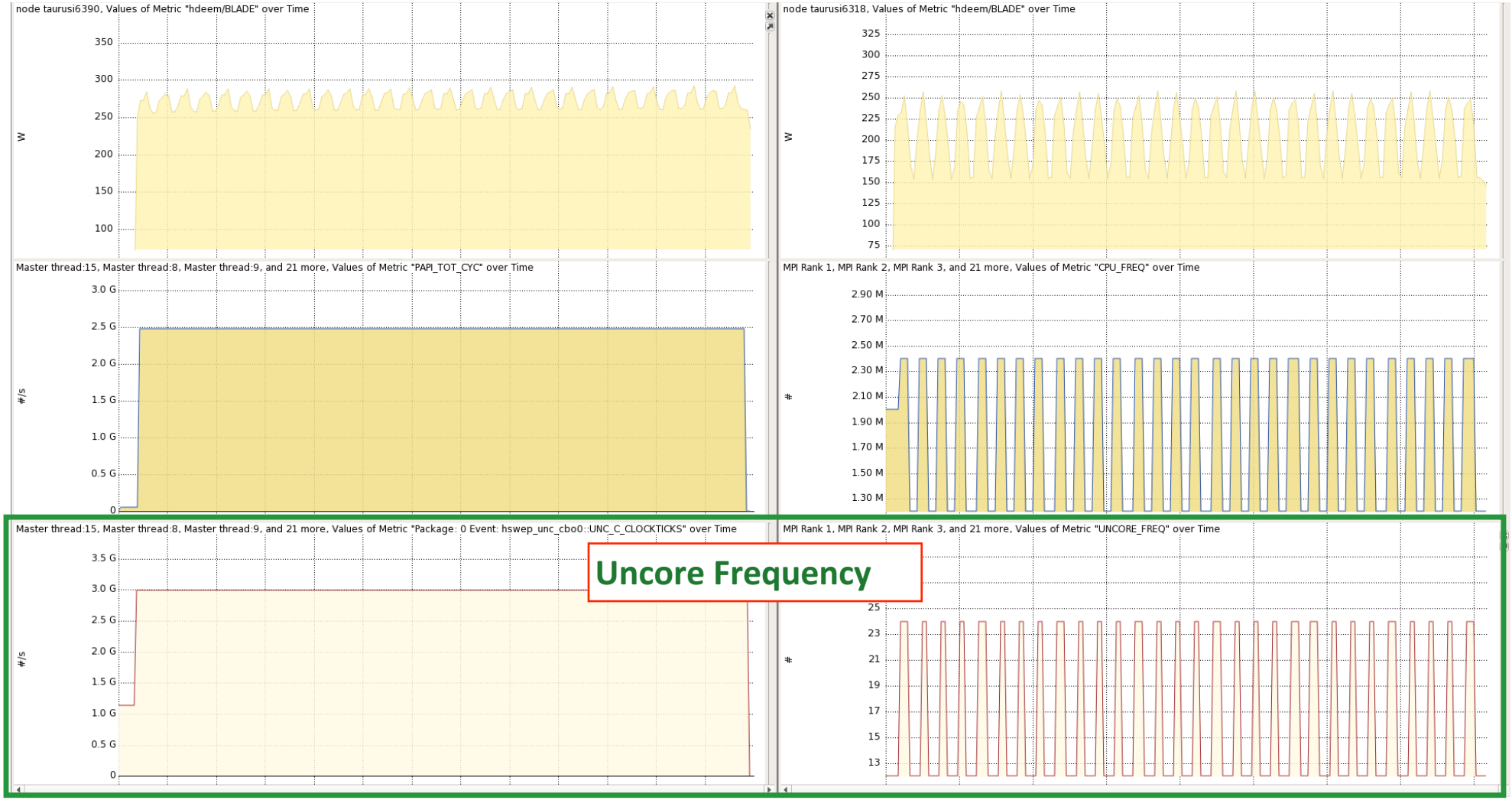
4.2. Results


Blade Power Consumption
node taurusi6390, Values of Metric "hdeem/BLADE" over Time
W
350
300
250
200
150
100
node taurusi6318, Values of Metric "hdeem/BLADE" over Time
W
325
300
275
250
225
200
175
150
125
100
75
Master thread:15, Master thread:8, Master thread:9, and 21 more, Values of Metric "PAPI_TOT_CYC" over Time
#/s
3.0 G
2.5 G
2.0 G
1.5 G
1.0 G
0.5 G
0
MPI Rank 1, MPI Rank 2, MPI Rank 3, and 21 more, Values of Metric "CPU_FREQ" over Time
#
2.90 M
2.70 M
2.50 M
2.30 M
2.10 M
1.90 M
1.70 M
1.50 M
1.30 M
Master thread:15, Master thread:8, Master thread:9, and 21 more, Values of Metric "Package: 0 Event: hswep_unc_cbo0::UNC_C_CLOCKTICKS" over Time
#/s
3.5 G
3.0 G
2.5 G
2.0 G
1.5 G
1.0 G
0.5 G
0
MPI Rank 1, MPI Rank 2, MPI Rank 3, and 21 more, Values of Metric "UNCORE_FREQ" over Time
#
29
27
25
23
21
19
17
15
13

4.2. Results


node taurusi6390, Values of Metric "hdeem/BLADE" over Time
W
350
300
250
200
150
100
node taurusi6318, Values of Metric "hdeem/BLADE" over Time
W
325
300
275
250
225
200
175
150
125
100
75
Master thread:15, Master thread:8, Master thread:9, and 21 more, Values of Metric "PAPI_TOT_CYC" over Time
#/s
3.0 G
2.5 G
2.0 G
1.5 G
1.0 G
0.5 G
0
CPU Core Frequency
MPI Rank 1, MPI Rank 2, MPI Rank 3, and 21 more, Values of Metric "CPU_FREQ" over Time
#
2.50 M
2.30 M
2.10 M
1.90 M
1.70 M
1.50 M
1.30 M
Master thread:15, Master thread:8, Master thread:9, and 21 more, Values of Metric "Package: 0 Event: hswep_unc_cbo0::UNC_C_CLOCKTICKS" over Time
#/s
3.5 G
3.0 G
2.5 G
2.0 G
1.5 G
1.0 G
0.5 G
0
MPI Rank 1, MPI Rank 2, MPI Rank 3, and 21 more, Values of Metric "UNCORE_FREQ" over Time
#
29
27
25
23
21
19
17
15
13

4.2. Results


node taurusi6390, Values of Metric "hdeem/BLADE" over Time
W
350
300
250
200
150
100
node taurusi6318, Values of Metric "hdeem/BLADE" over Time
W
325
300
275
250
225
200
175
150
125
100
75
Master thread:15, Master thread:8, Master thread:9, and 21 more, Values of Metric "PAPI_TOT_CYC" over Time
#/s
3.0 G
2.5 G
2.0 G
1.5 G
1.0 G
0.5 G
0
MPI Rank 1, MPI Rank 2, MPI Rank 3, and 21 more, Values of Metric "CPU_FREQ" over Time
#
2.90 M
2.70 M
2.50 M
2.30 M
2.10 M
1.90 M
1.70 M
1.50 M
1.30 M
Master thread:15, Master thread:8, Master thread:9, and 21 more, Values of Metric "Package: 0 Event: hswep_unc_cbo0::UNC_C_CLOCKTICKS" over Time
#/s
3.5 G
3.0 G
2.5 G
2.0 G
1.5 G
1.0 G
0.5 G
0
Uncore Frequency
MPI Rank 1, MPI Rank 2, MPI Rank 3, and 21 more, Values of Metric "UNCORE_FREQ" over Time
#
25
23
21
19
17
15
13

5 Summary

5.1. Next steps

5.2. Contact details

6.1. Next steps

1. Performed experiments on the Taurus cluster at TU Dresden on 5-20 nodes.
   - Larger experiments (~100 nodes) with larger application input on Taurus.

2. Obtained preparatory access (Type B) on MareNostrum4 cluster.
   - Install READEX on MN4 and run experiments.

3. ICHEC and BSC implementing Alya-specific application parameter tuning.

6.2. Contact details

Venkatesh Kannan

Irish Centre for High-End Computing (ICHEC)

venkatesh.kannan@ichec.ie

READEX website

www.readex.eu

Alya website

https://www.bsc.es/es/computer-applications/alya-system

THANK YOU FOR YOUR ATTENTION

www.prace-ri.eu

PRACE T7.2-5IP

Towards Energy-efficient Exascale Computing: A Use-case Applying READEX to Alya

PRACE @ ISC 2018

Ramón Martínez Carreras
Irish Centre for High-End Computing (ICHEC)

ramon.carreras@ichec.ie