Power-aware HPC on Intel Xeon Phi KNL Processors

Azzam Haidar, Heike Jagode, Asim YarKhan, Phil Vaccaro,
Stan Tomov, and Jack Dongarra

Innovative Computing Laboratory
Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville

ISC High Performance 2017 (ISC'17)
Frankfurt, Germany
June 18 – 22, 2017

Intel Parallel Computing Centers

INNOVATIVE COMPUTING LABORATORY
THE UNIVERSITY of TENNESSEE

Outline

- How to design power and energy-aware numerical libraries ?
- PAPI-based tools to quantify, understand and control power usage (through power capping)
- Characterize algorithms (from BLAS and benchmarks) and quantify possible benefits from power capping on Intel KNL processors
- Developments can be used to automatically analyze and build into libraries power & energy control based on feedback from computation (at runtime)

PAPI

- Middleware that provides a **consistent interface** and methodology for the performance counter hardware found in most major microprocessors
- PAPI enables software engineers to see, in near real time, the relation between **SW performance** and **HW events**

SUPPORTED ARCHITECTURES:

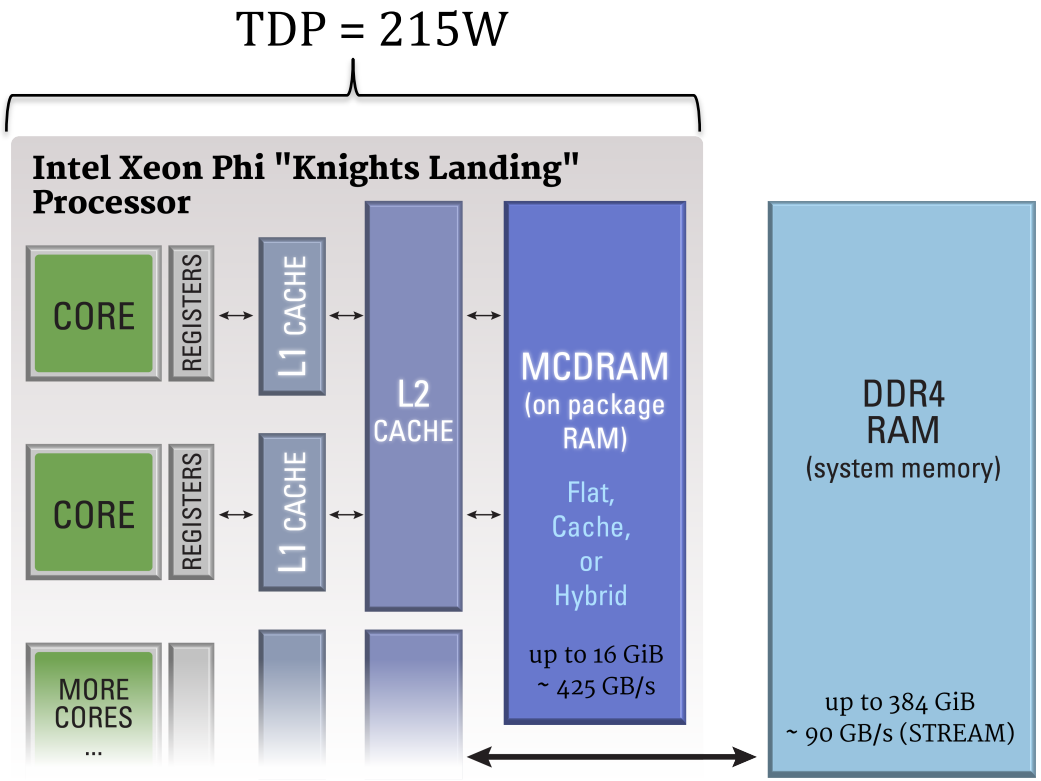
- AMD
- CRAY: Aris, Gemini, power
- IBM Blue Gene Series, Q: 5D-Torus, I/O system, CNK, EMON power/energy
- IBM Power Series
- Intel Westmere, Sandy|Ivy Bridge, Haswell, Broadwell, **Skylake**, Knights Corner | **Landing**
- ARM Cortex A8, A9, A15**, ARM64**
- NVidia Tesla, Kepler, NVML: **CUDA support for multiple GPUs; PC Sampling**
- Infiniband
- Intel RAPL (power/energy); **power capping**
- Intel KNC**, KNL power/energy**

COMPONENT PAPI:

- provides access to a collection of components that expose performance measurement opportunities across the system as a whole, including network, the I/O system, the Compute Node Kernel, power/ energy

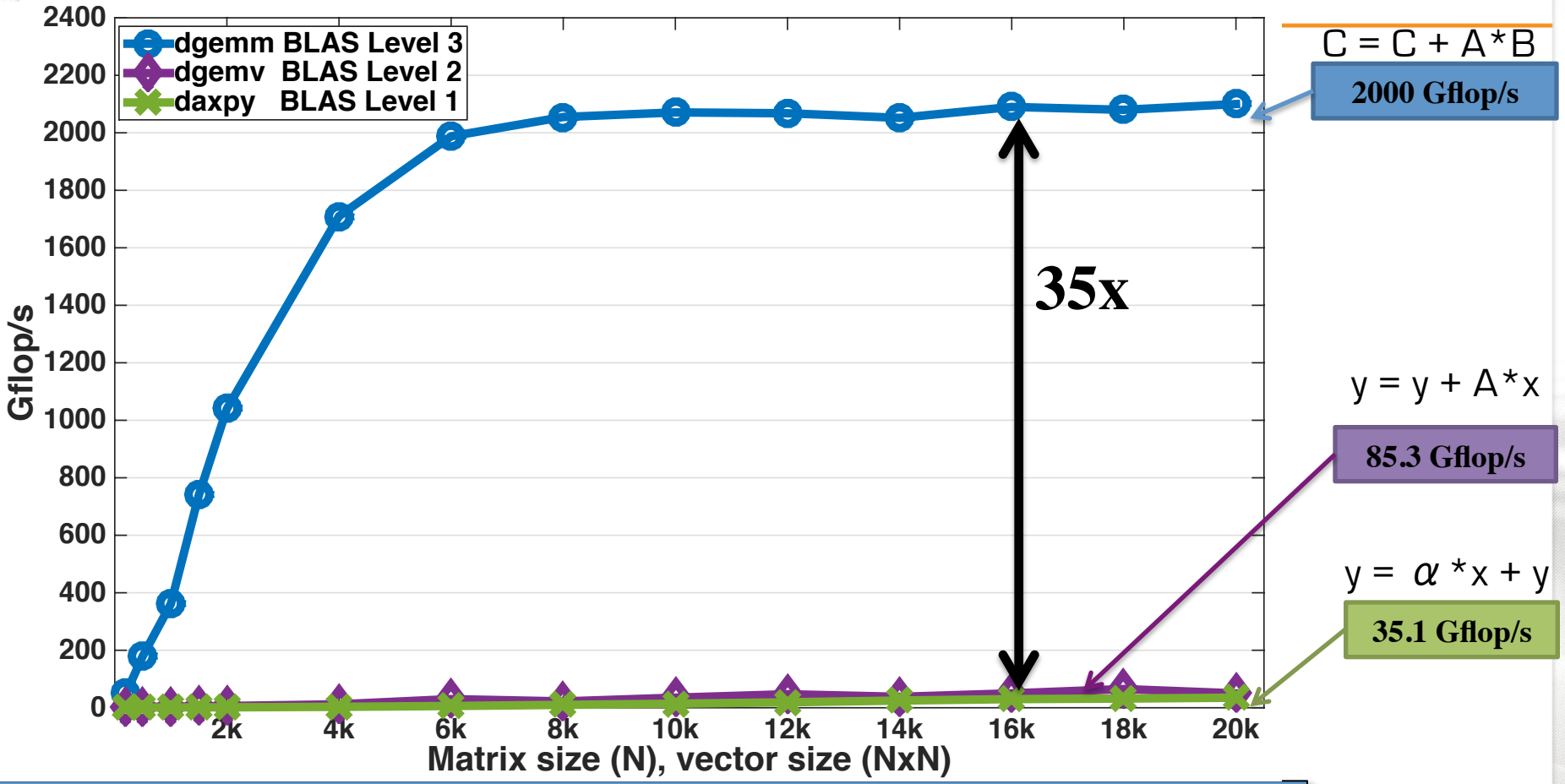
Intel Knights Landing: FLAT mode

- KNL is in FLAT mode (where entire MCDRAM is used as addressable memory)
  → memory allocations are treated similarly to DDR4 memory allocations
- KNL Thermal Design Point (TDP) is 215W (for main SKUs):



Level 1, 2 and 3 BLAS

68 cores **Intel Xeon Phi KNL**, 1.3 GHz, Peak DP = 2662 Gflop/s



68 cores Intel Xeon Phi KNL, 1.3 GHz
The theoretical peak double precision is 2662 Gflop/s
Compiled with icc and using Intel MKL 2017b1 20160506

PAPI for power-aware computing

- We use PAPI's latest **powercap component** for measurement, control, and performance analysis
  - PAPI power components in the past supported **only reading** power information
  - New component exposes RAPL functionality to allow users to **read and write** power

PAPI for power-aware computing

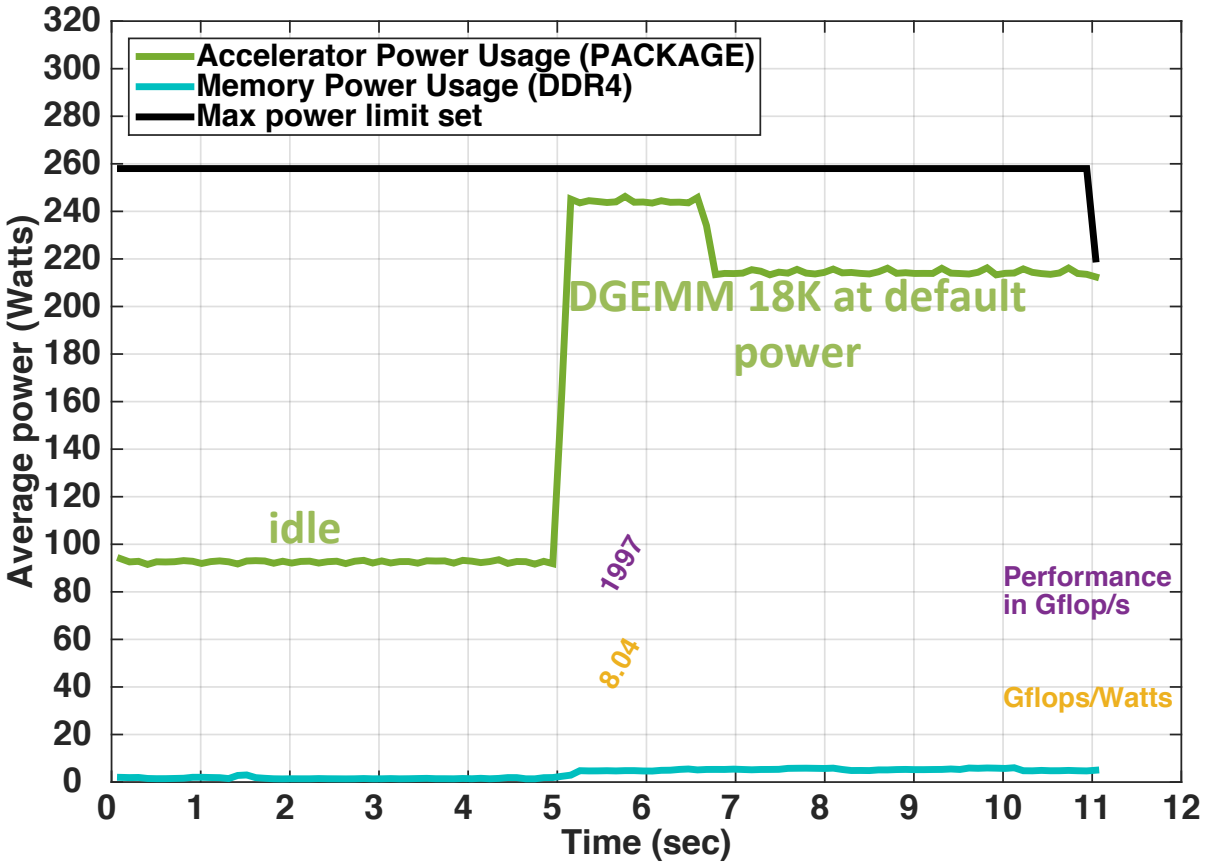
- We use PAPI's latest **powercap component** for measurement, control, and performance analysis
  - PAPI power components in the past supported **only reading** power information
  - New component exposes RAPL functionality to allow users to **read and write** power

- Study numerical building blocks of varying computational intensity
- Use PAPI powercap component to detect power optimization opportunities

→ **Objective:** Cap the power on the architecture to reduce power usage while keeping the execution time constant → **Energy Savings** !!!

Level 3 BLAS DGEMM on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

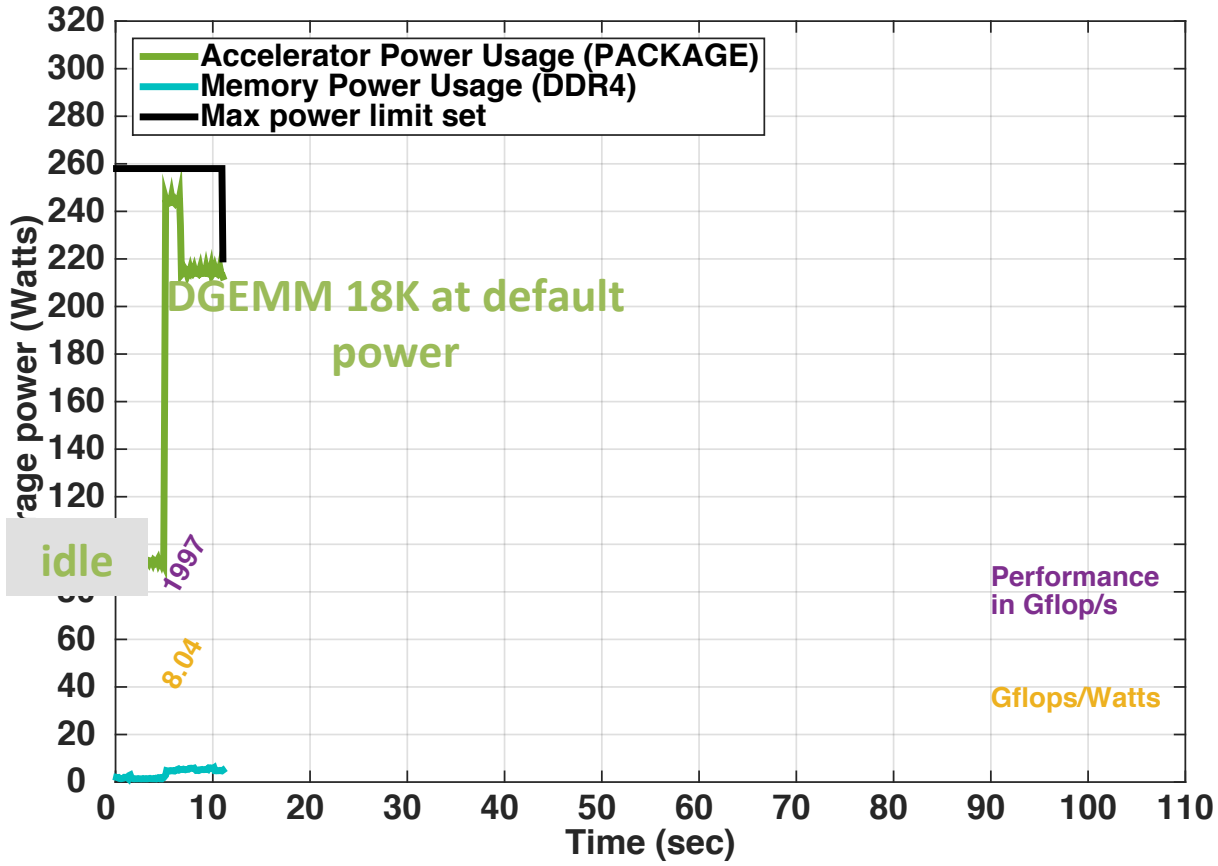


set pwr limit default
DGEMM size 18K

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s



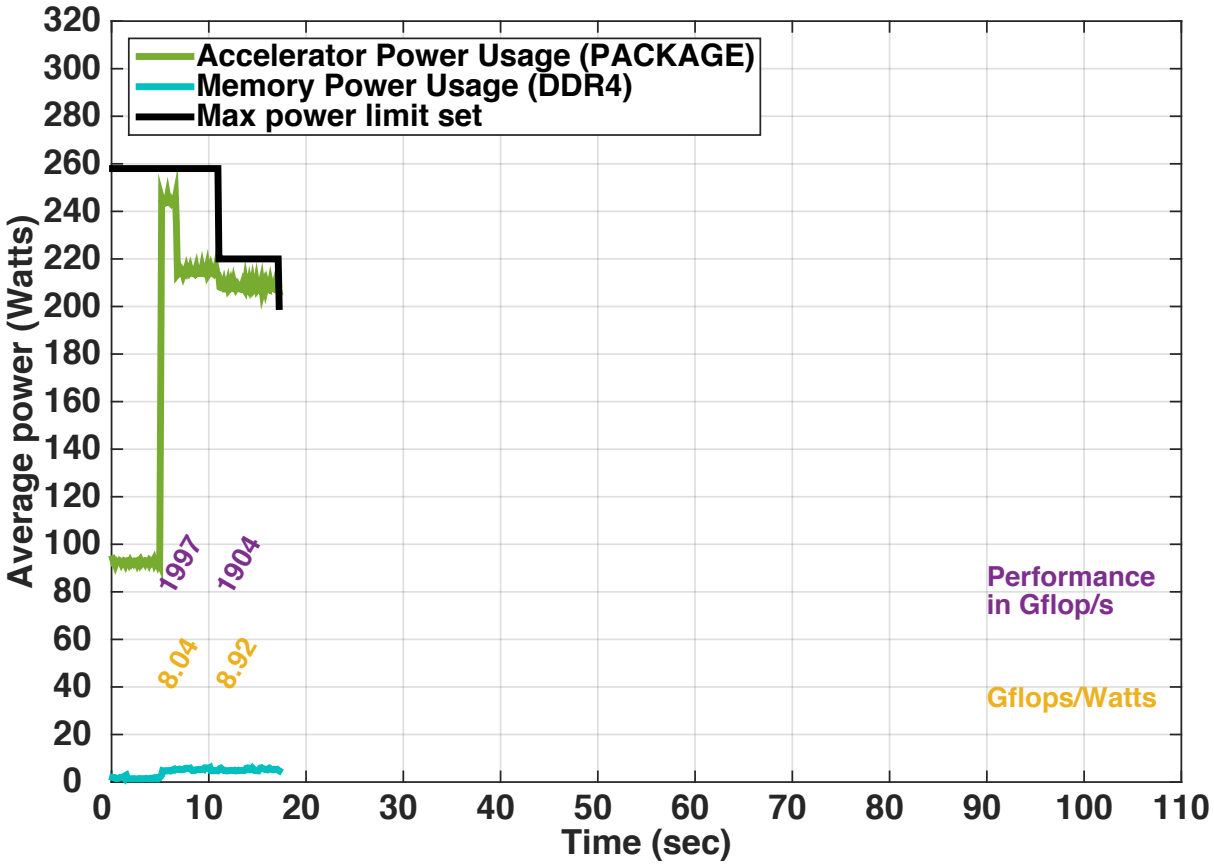
set pwr limit default

DGEMM size 18K

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

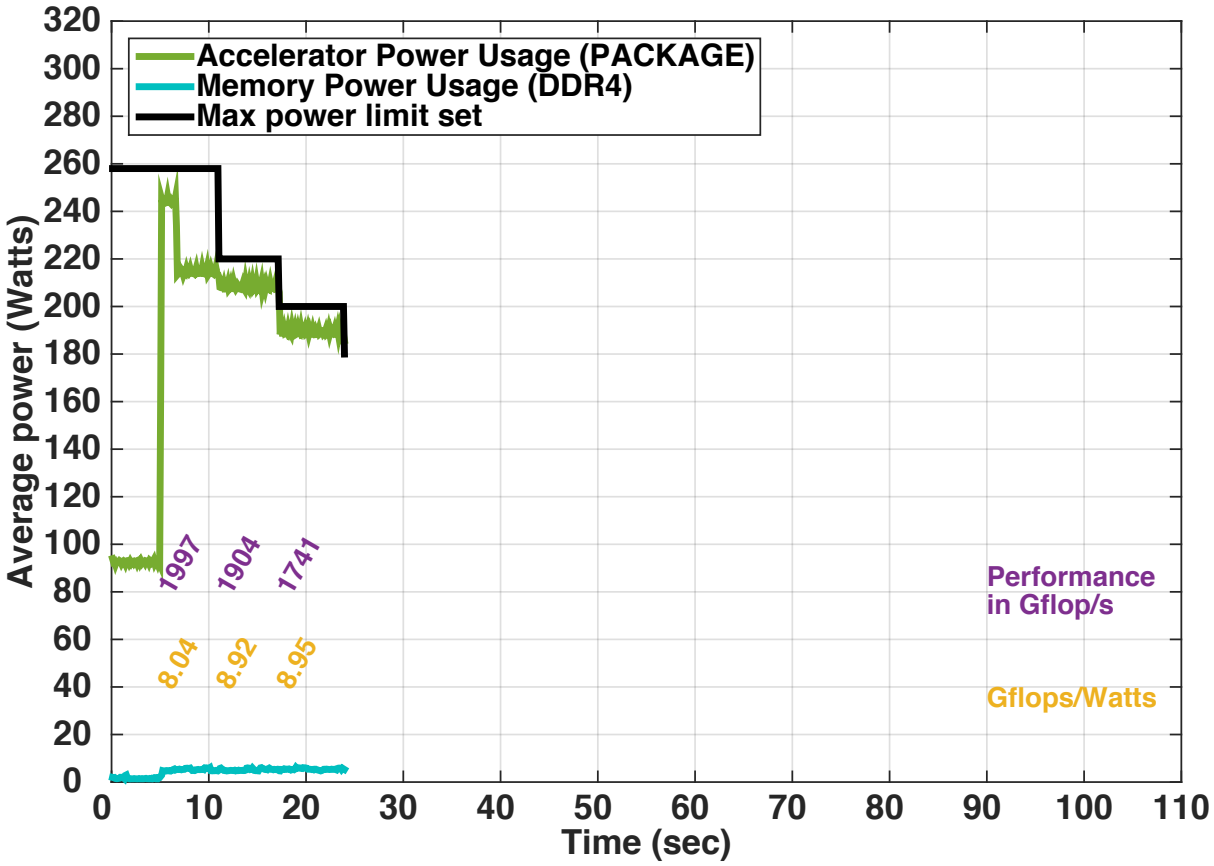


set pwr limit default
DGEMM size 18K
set pwr limit 220W
DGEMM size 18K

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

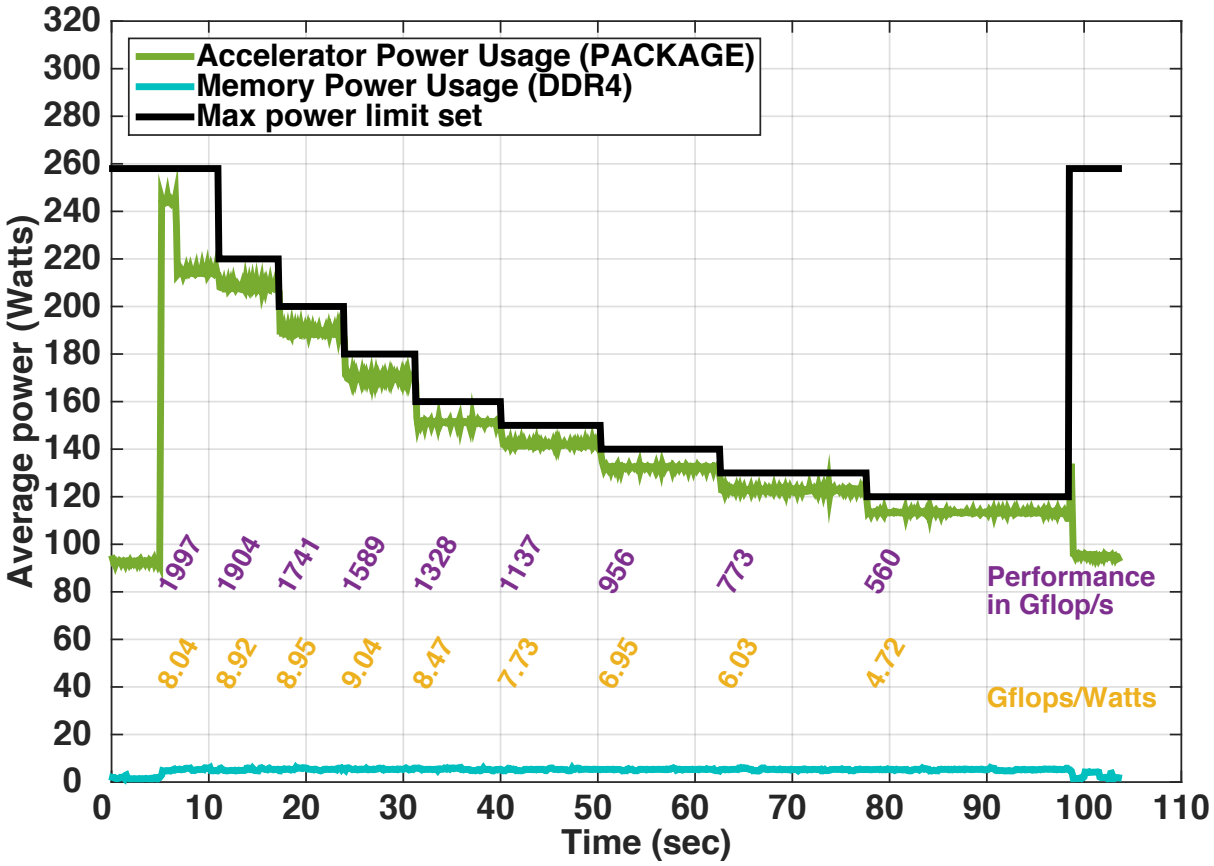


set pwr limit default
DGEMM size 18K
set pwr limit 220W
DGEMM size 18K
set pwr limit 200W
DGEMM size 18K

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

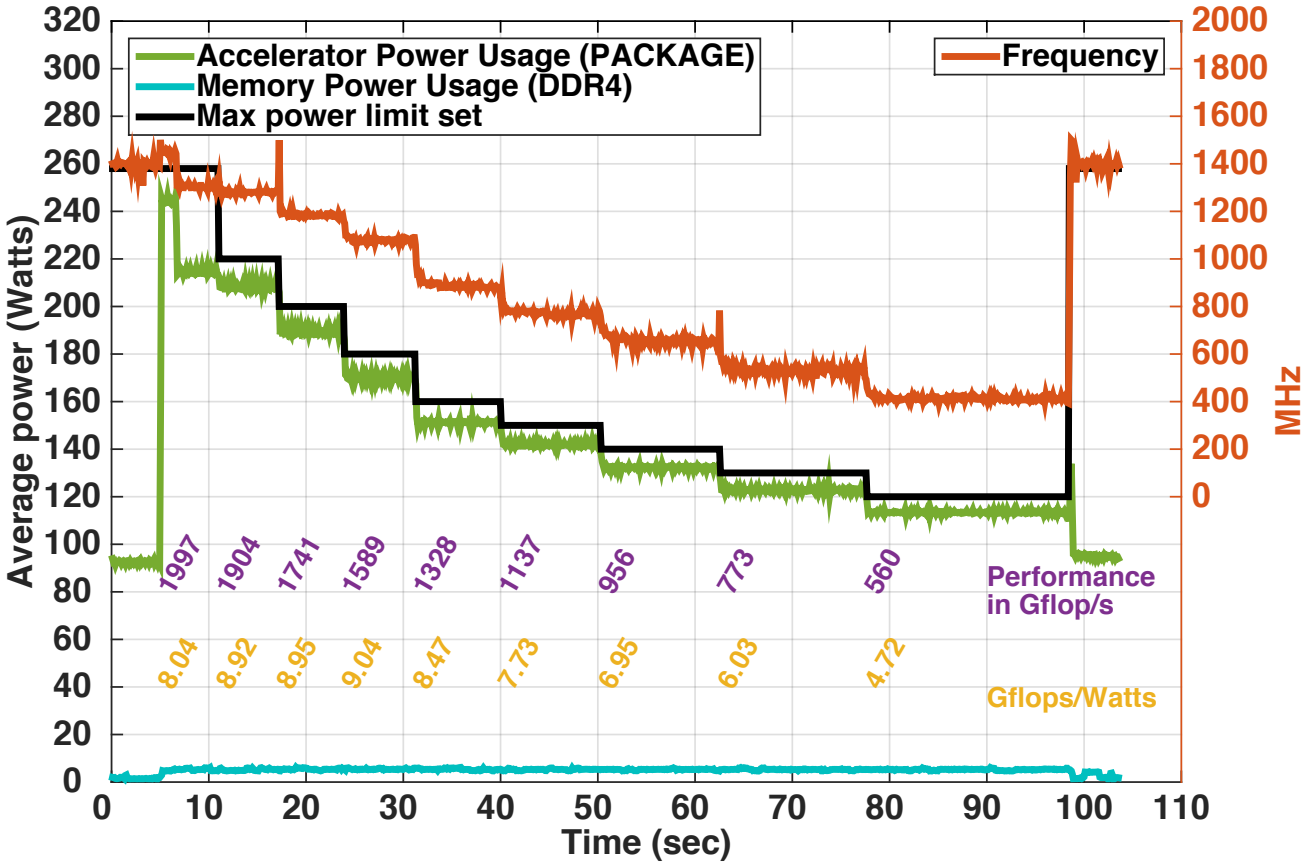


set pwr limit default
DGEMM size 18K
set pwr limit 220W
DGEMM size 18K
set pwr limit 200W
DGEMM size 18K
set pwr limit 180W
DGEMM size 18K
set pwr limit 160W
DGEMM size 18K
set pwr limit 150W
....

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

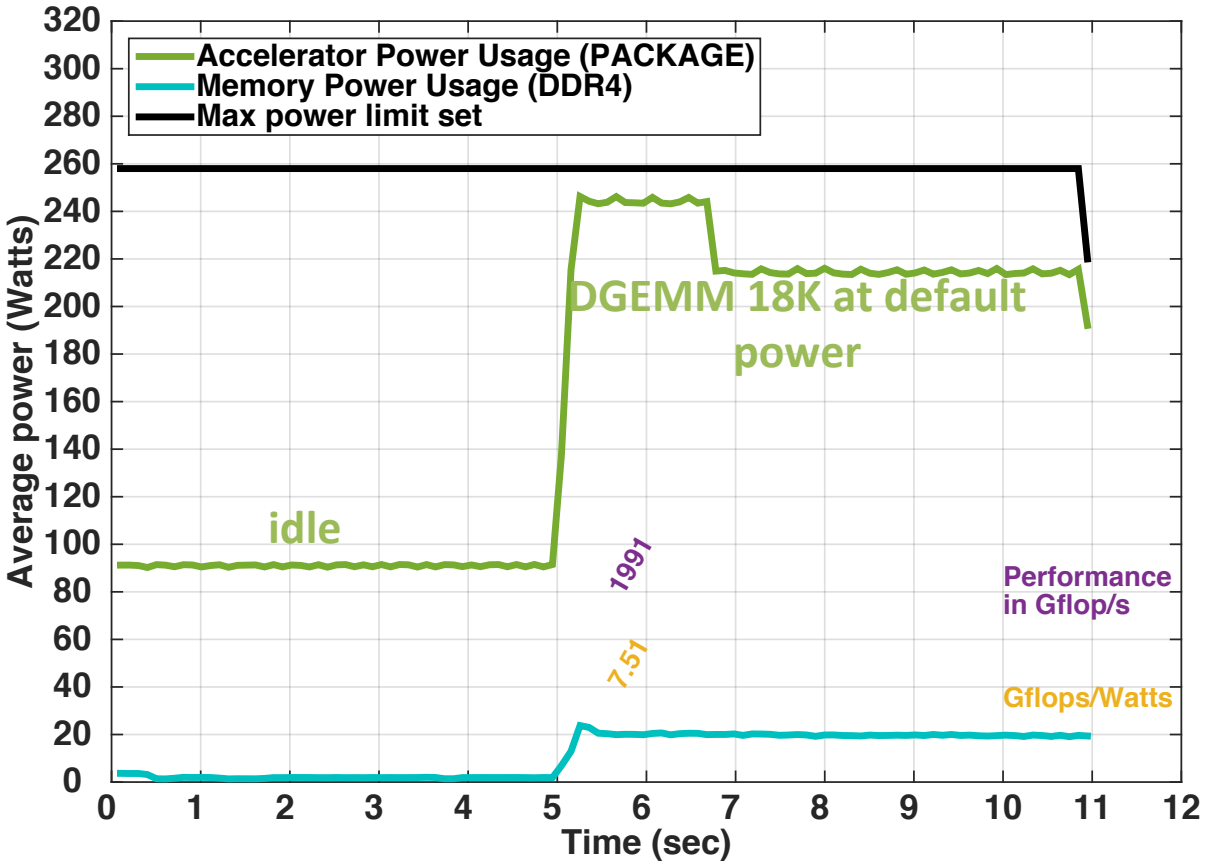


- When power cap kick-on the frequency is decreased which confirm that capping affect through the DVFS

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

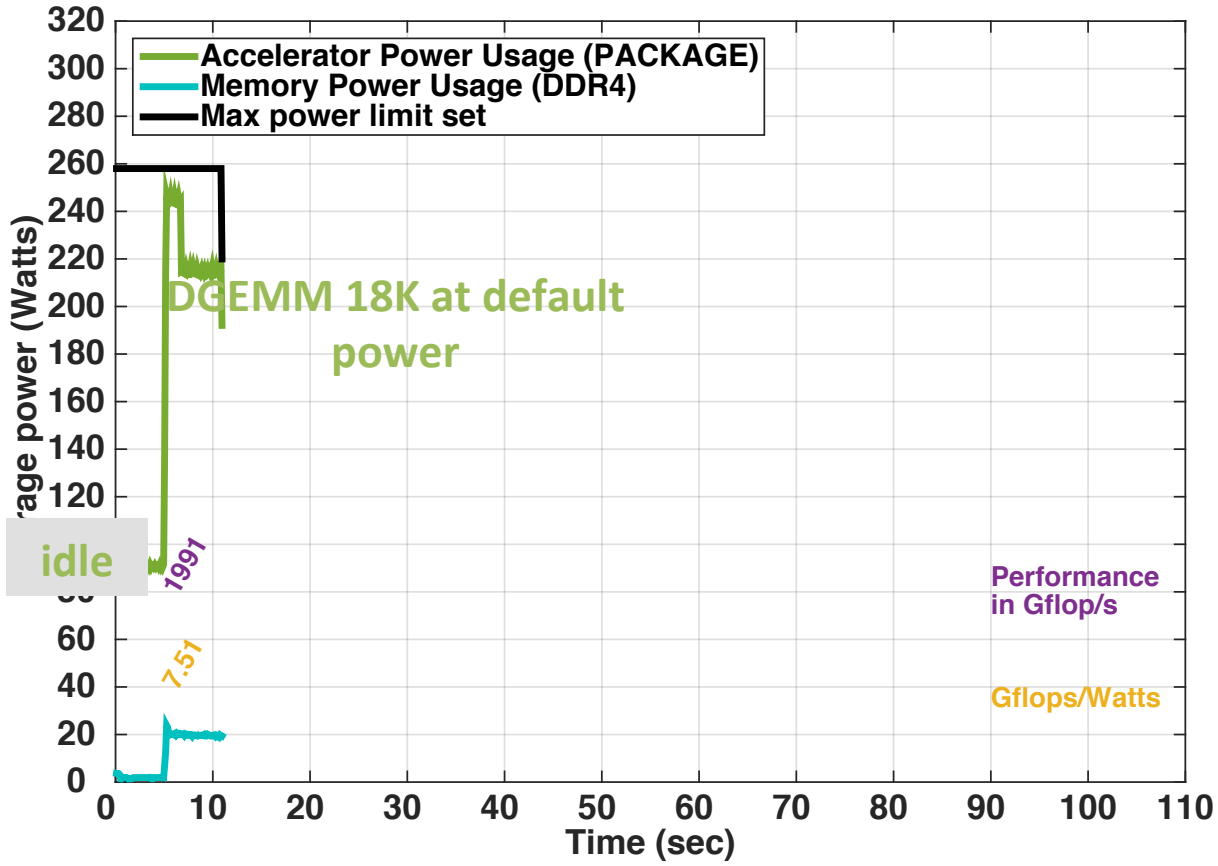


set pwr limit default
DGEMM size 18K

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s



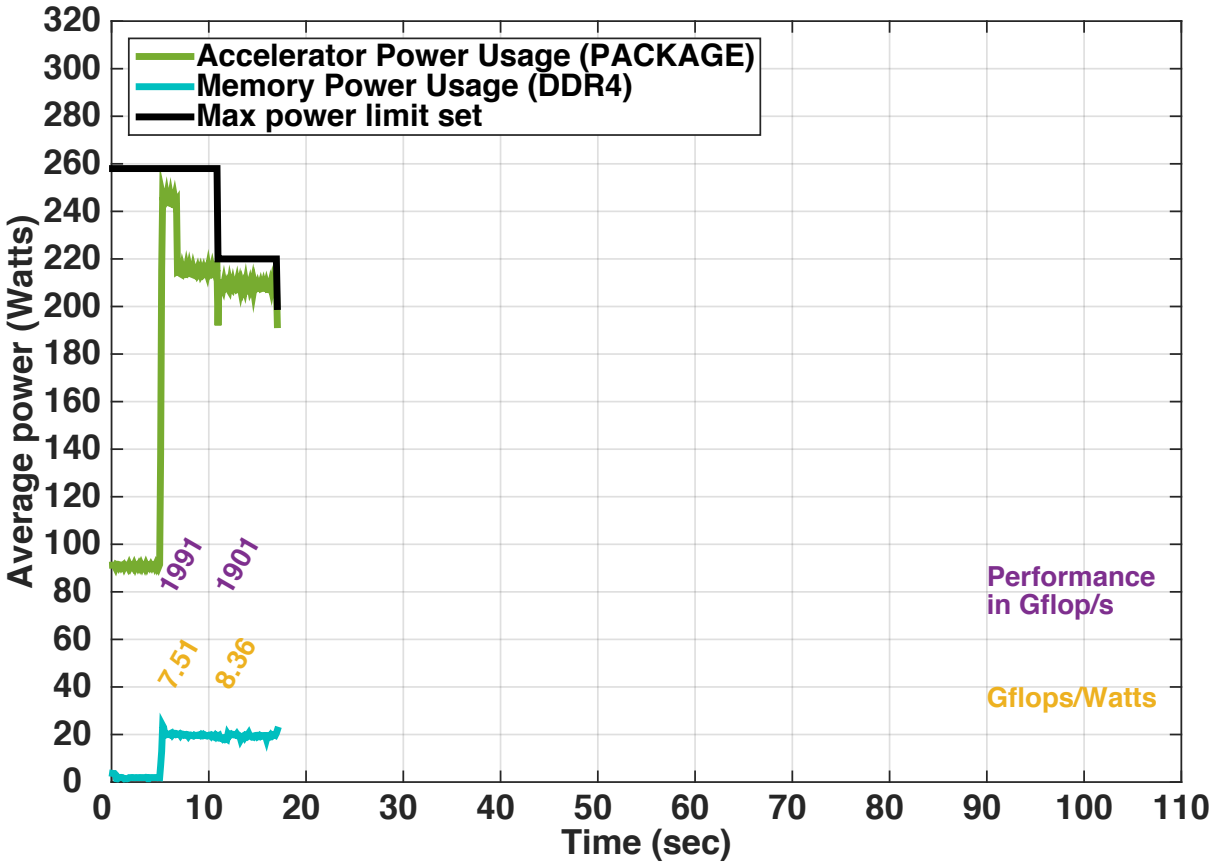
set pwr limit default

DGEMM size 18K

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

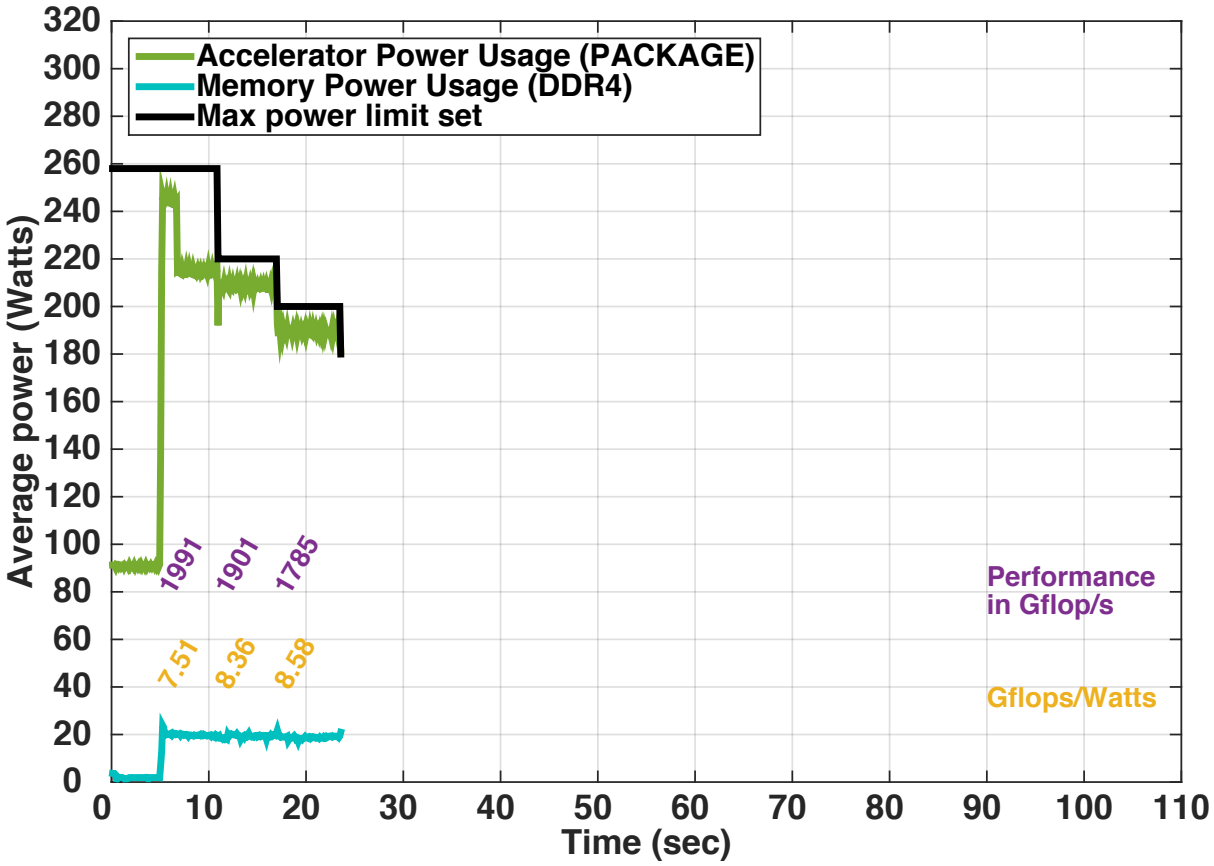


set pwr limit default
DGEMM size 18K
set pwr limit 220W
DGEMM size 18K

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

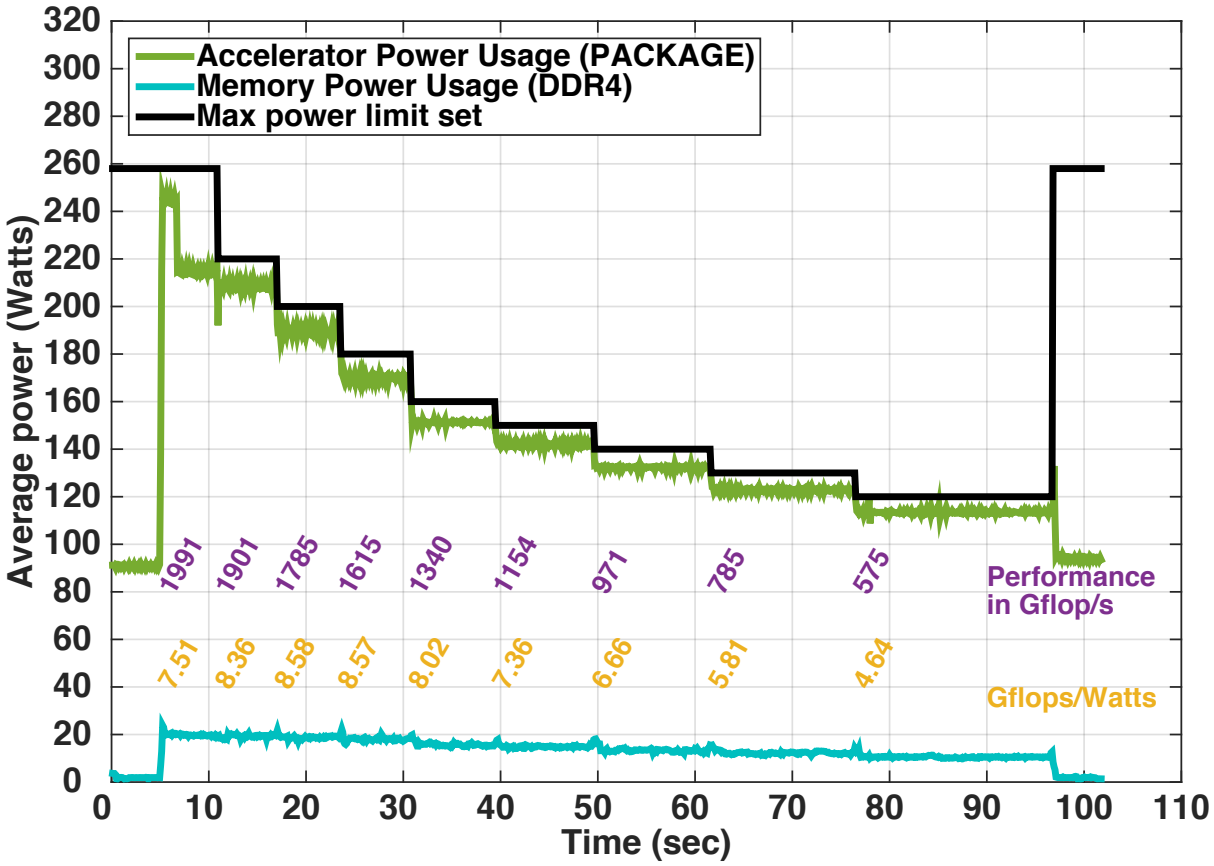


set pwr limit default
DGEMM size 18K
set pwr limit 220W
DGEMM size 18K
set pwr limit 200W
DGEMM size 18K

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

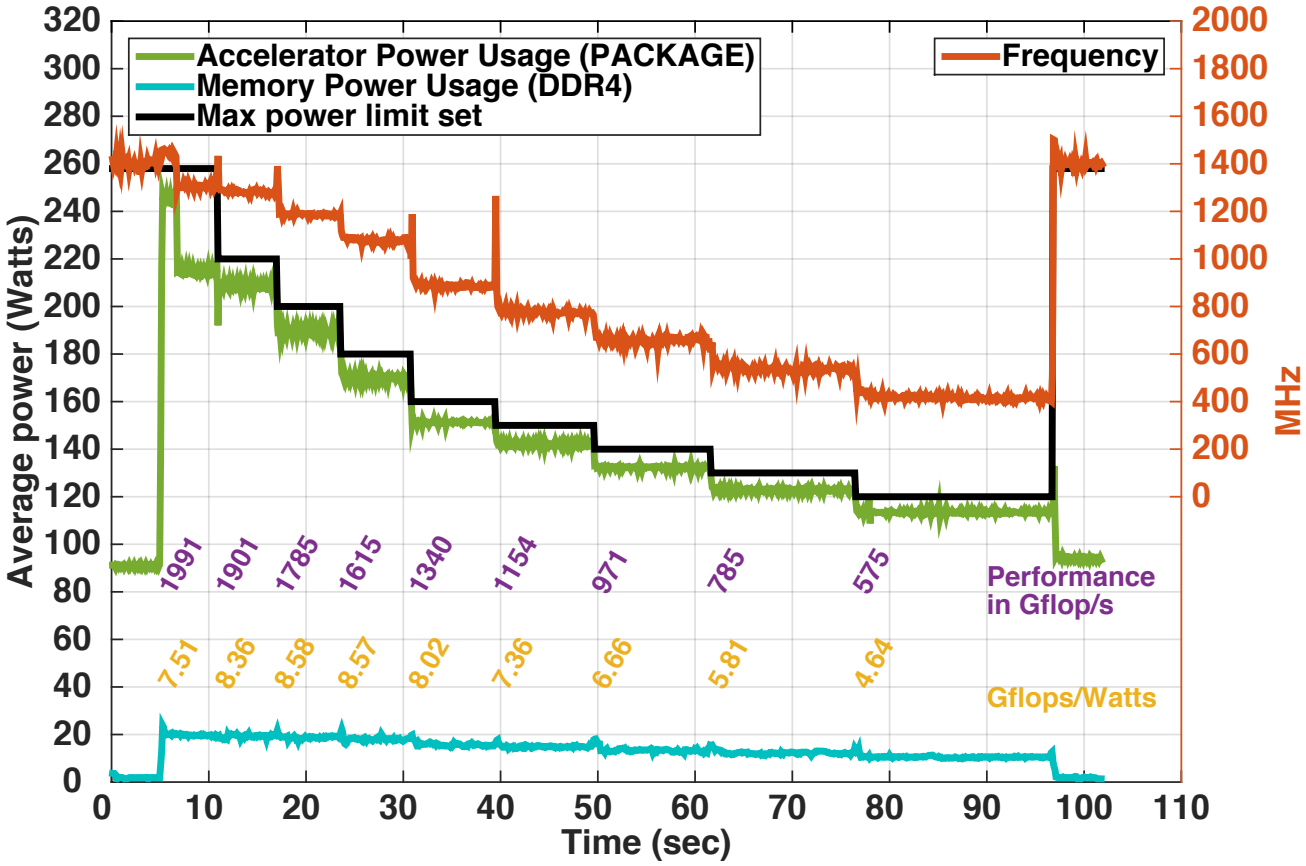


set pwr limit default
DGEMM size 18K
set pwr limit 220W
DGEMM size 18K
set pwr limit 200W
DGEMM size 18K
set pwr limit 180W
DGEMM size 18K
set pwr limit 160W
DGEMM size 18K
set pwr limit 150W
....

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

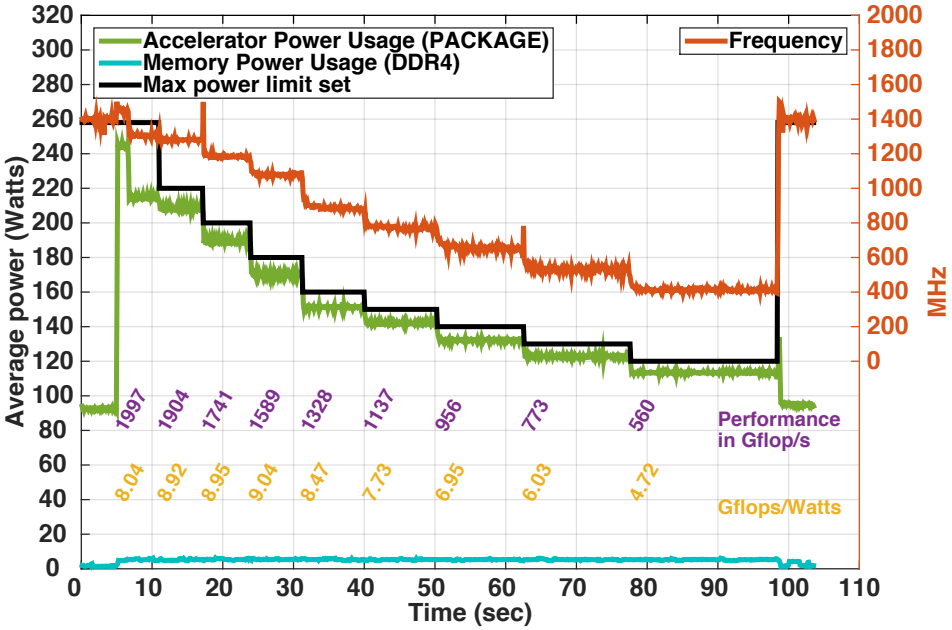


- When power cap kick-on the frequency is decreased which confirm that capping affect through the DVFS

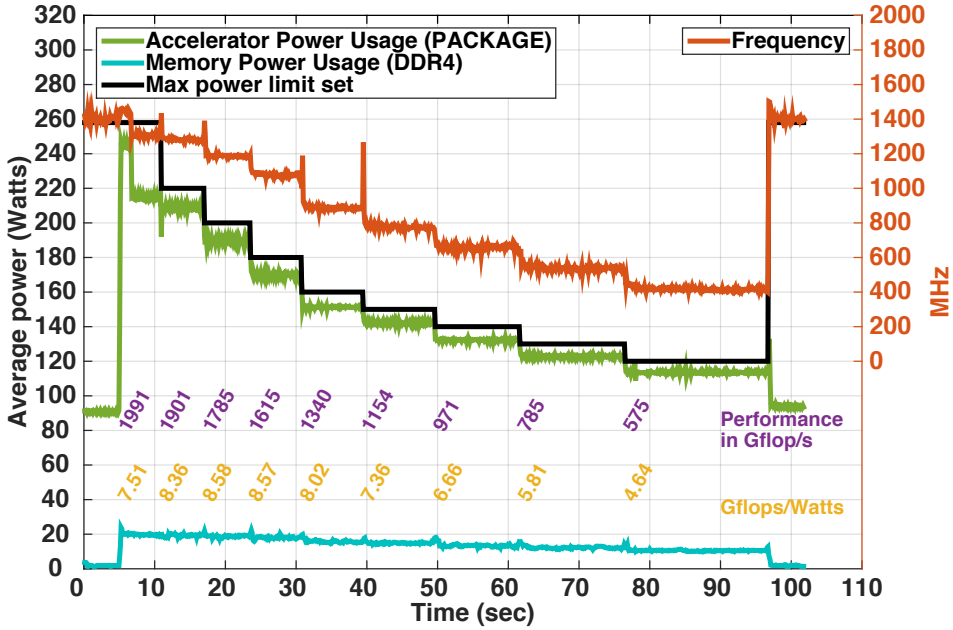
DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 3 BLAS DGEMM on KNL MCDRAM/DDR4

MCDRAM



DDR4



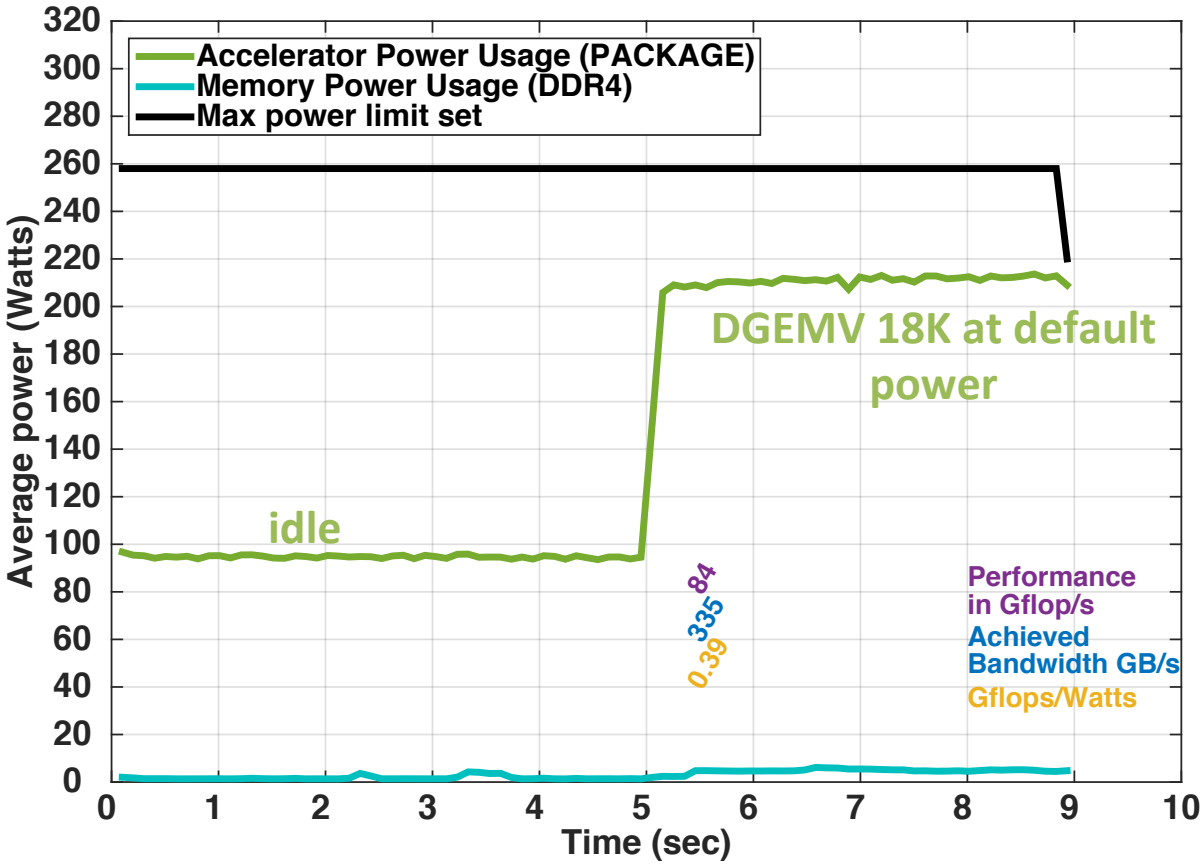
Lesson for DGEMM type of operations (compute intensive):

- **Capping can reduce performance, but sometimes can hold the energy efficiency (Gflos/Watts).**

DGEMM is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

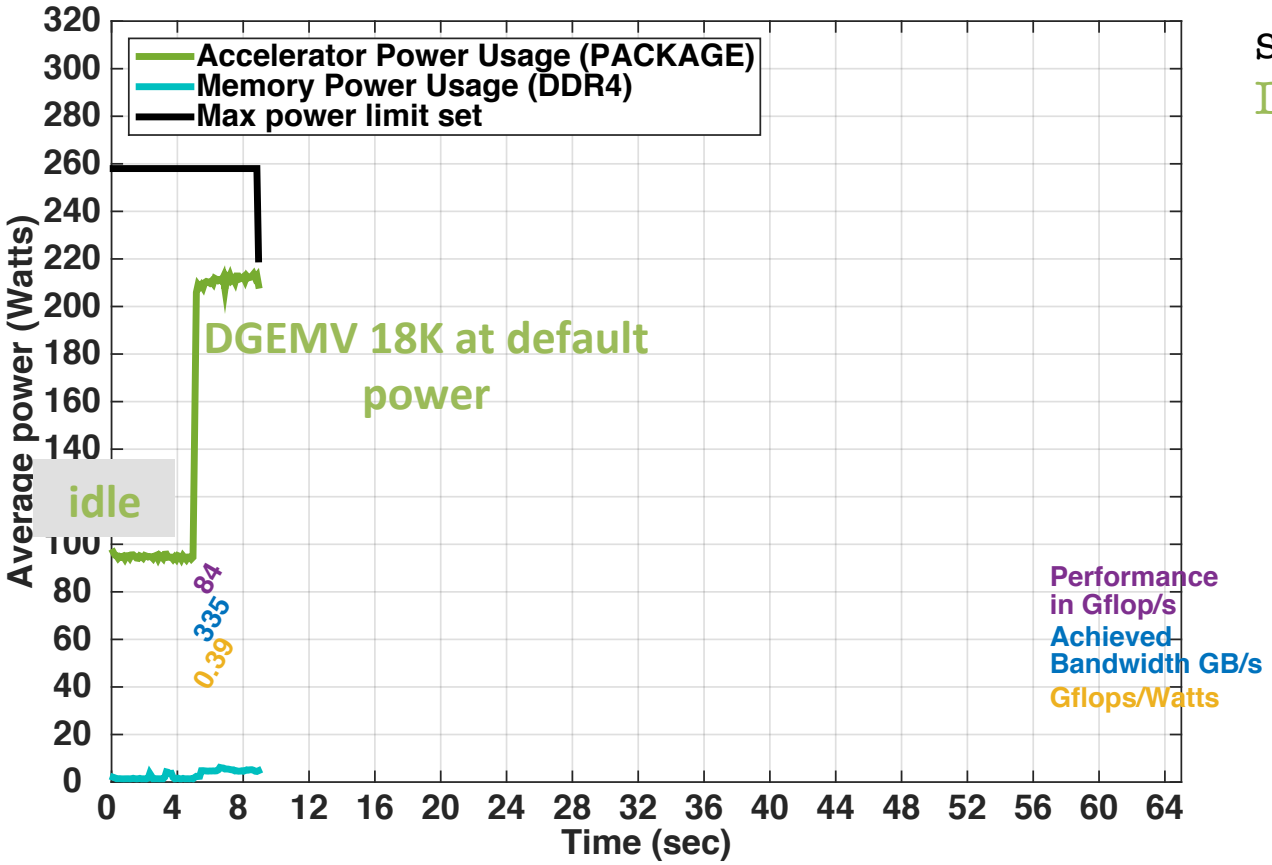


set pwr limit default
DGEMV size 18K

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s



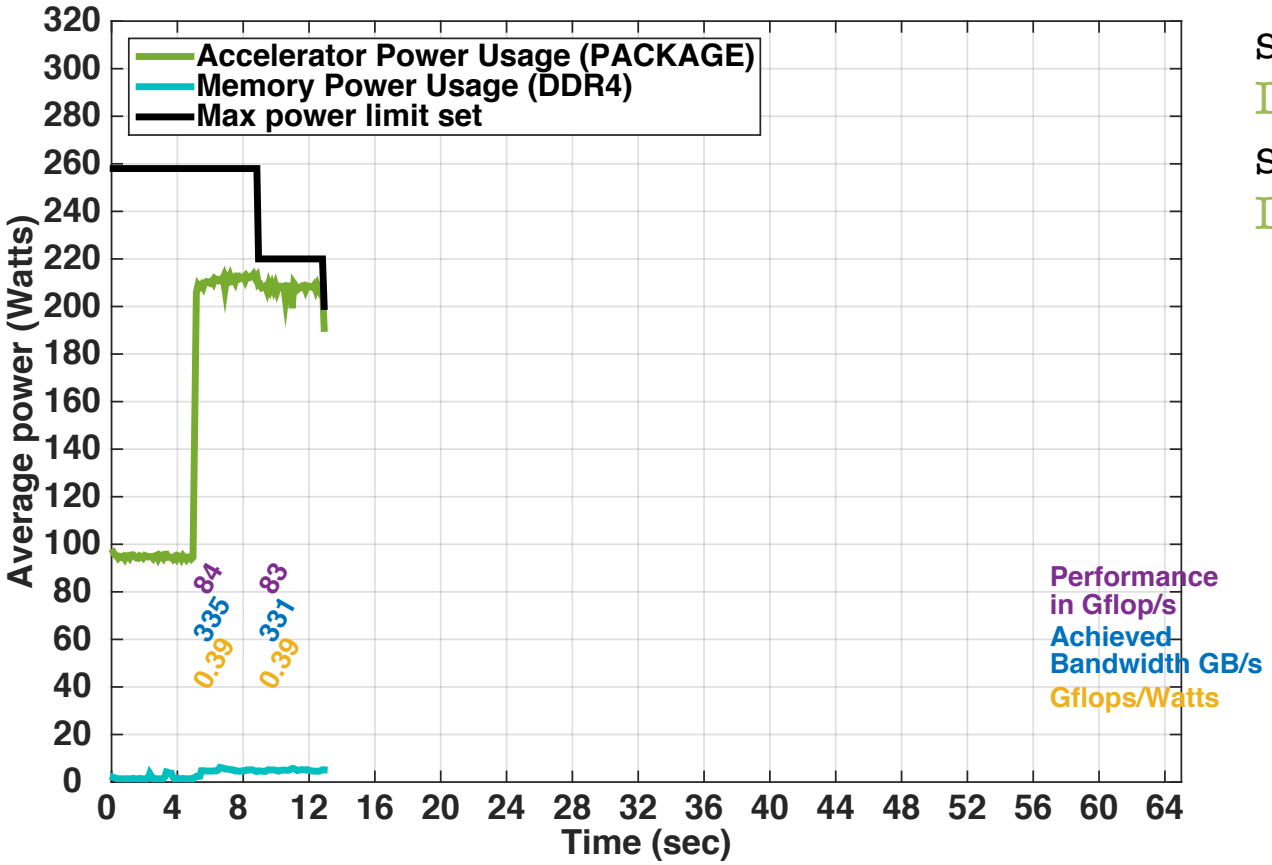
set pwr limit default

DGEMV size 18K

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

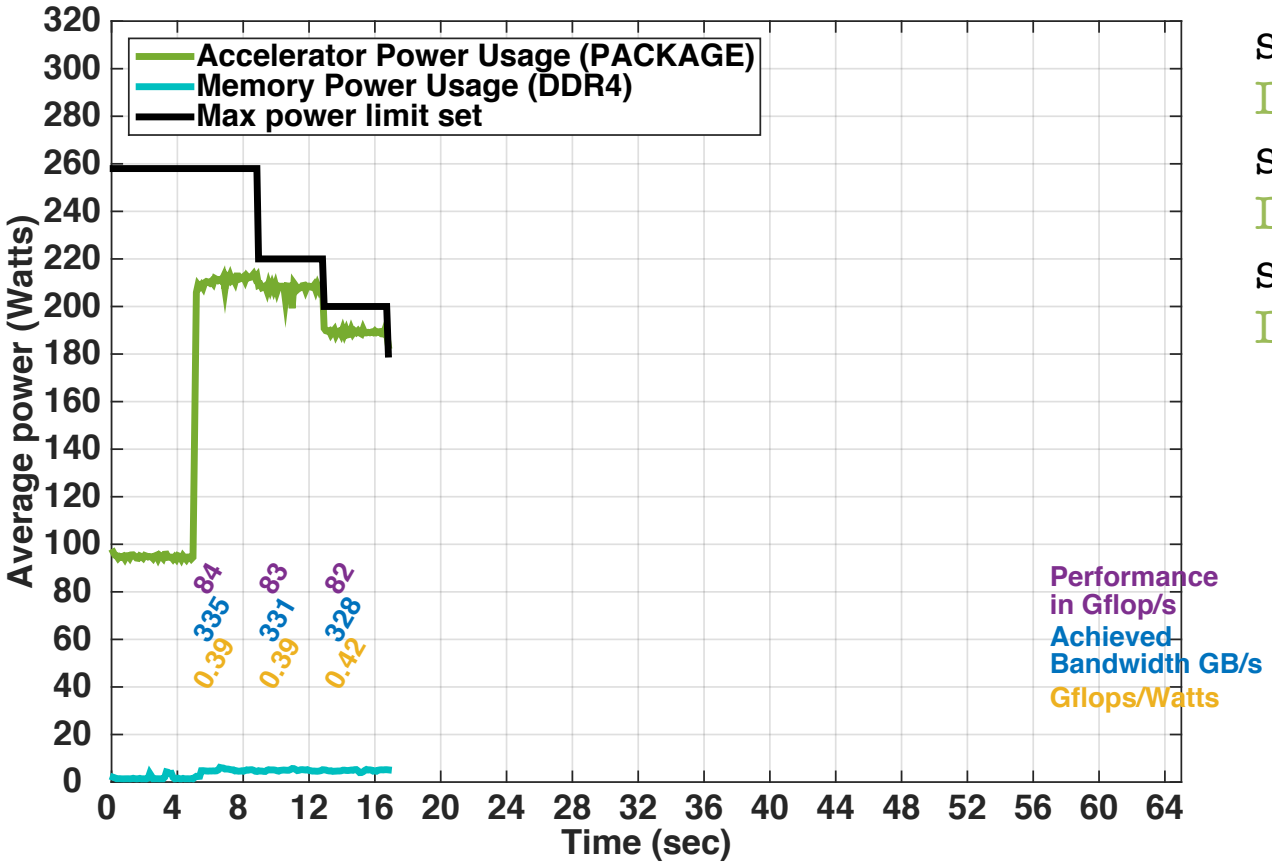


set pwr limit default
DGEMV size 18K
set pwr limit 220W
DGEMV size 18K

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

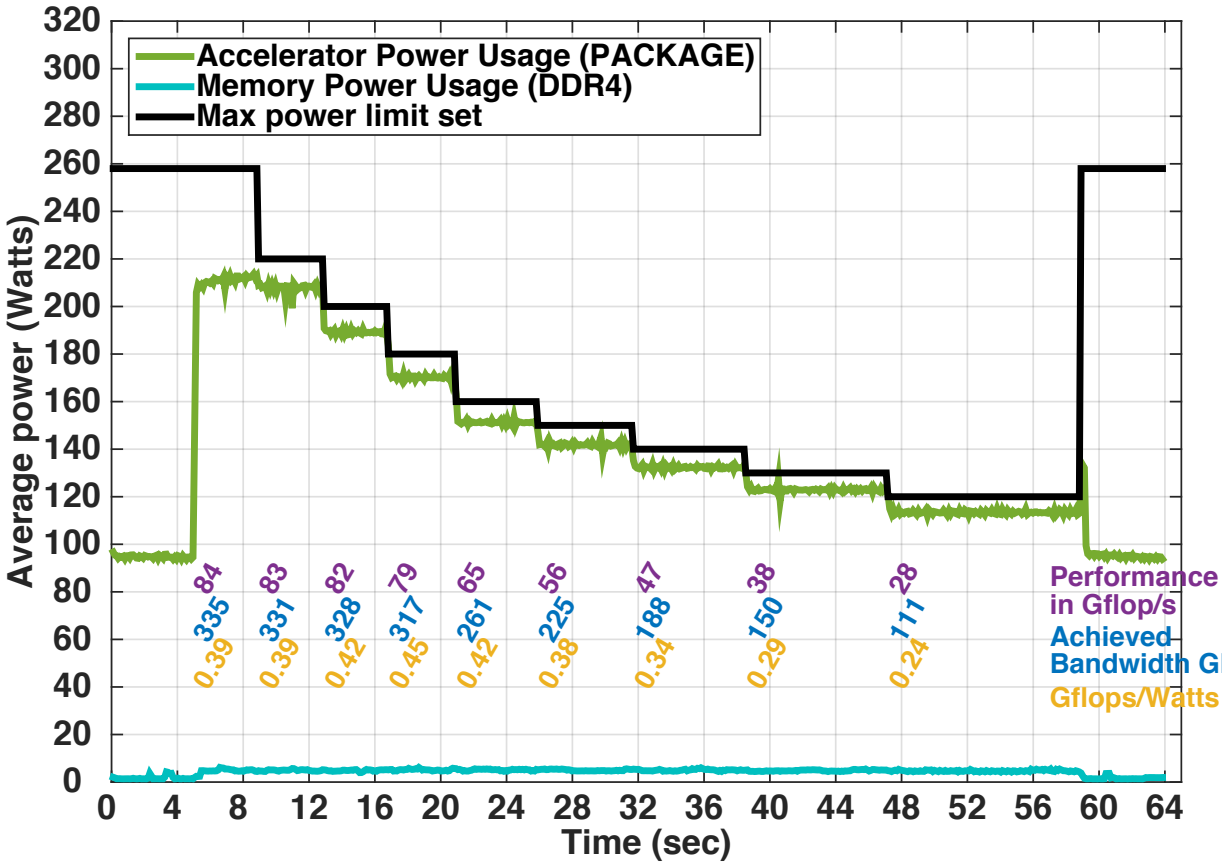


set pwr limit default
DGEMV size 18K
set pwr limit 220W
DGEMV size 18K
set pwr limit 200W
DGEMV size 18K

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

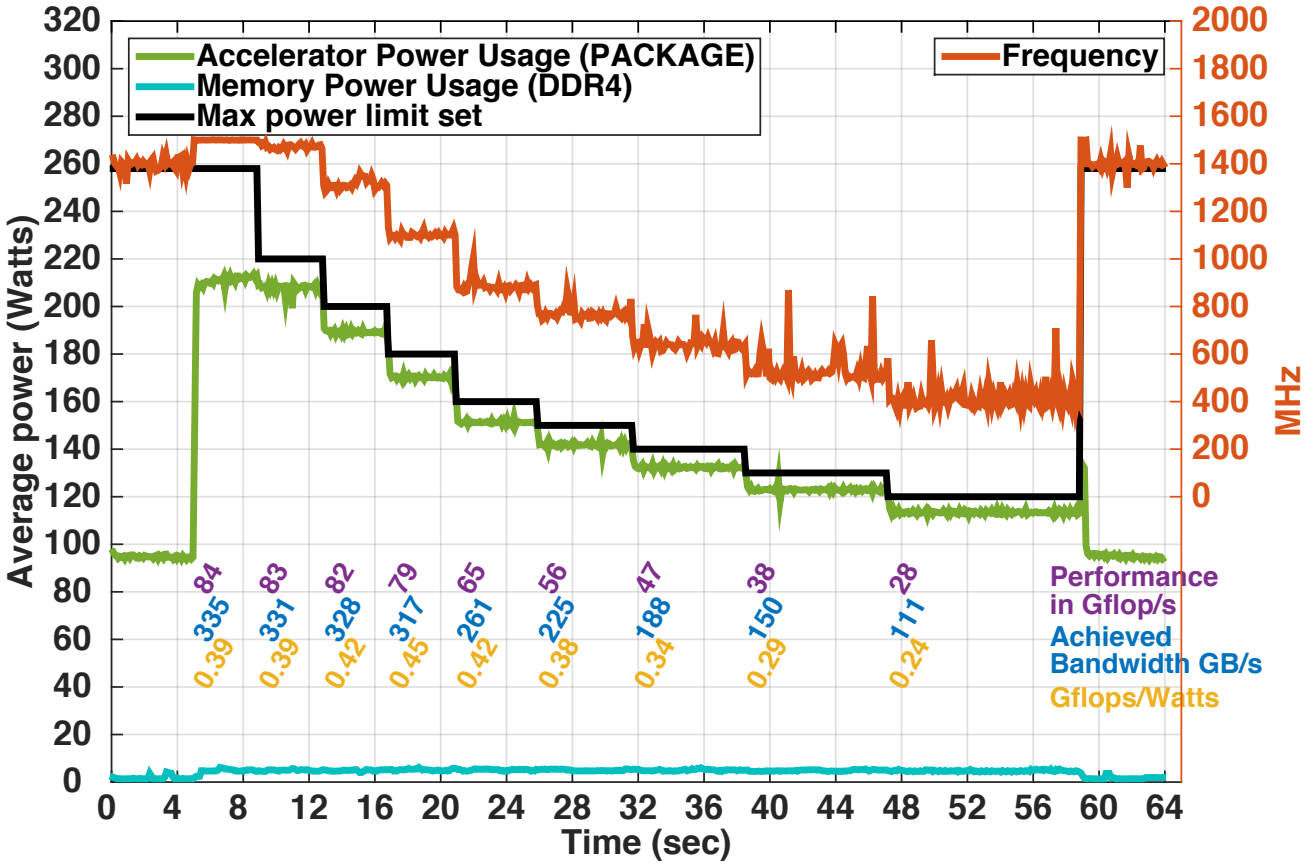


set pwr limit default
DGEMV size 18K
set pwr limit 220W
DGEMV size 18K
set pwr limit 200W
DGEMV size 18K
set pwr limit 180W
DGEMV size 18K
set pwr limit 160W
DGEMV size 18K
set pwr limit 150W
....

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

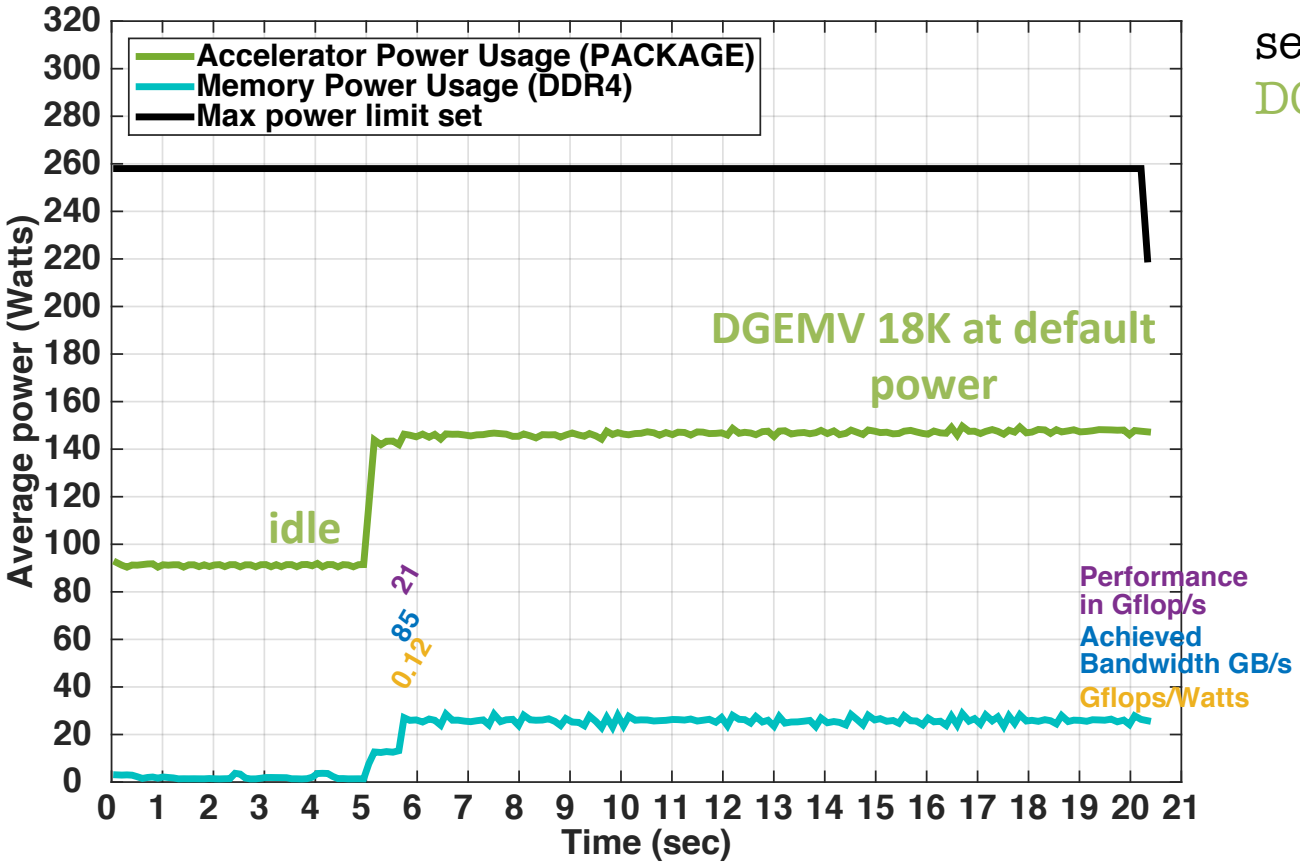


- **Frequency is not affected until the cap kick-on**

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s



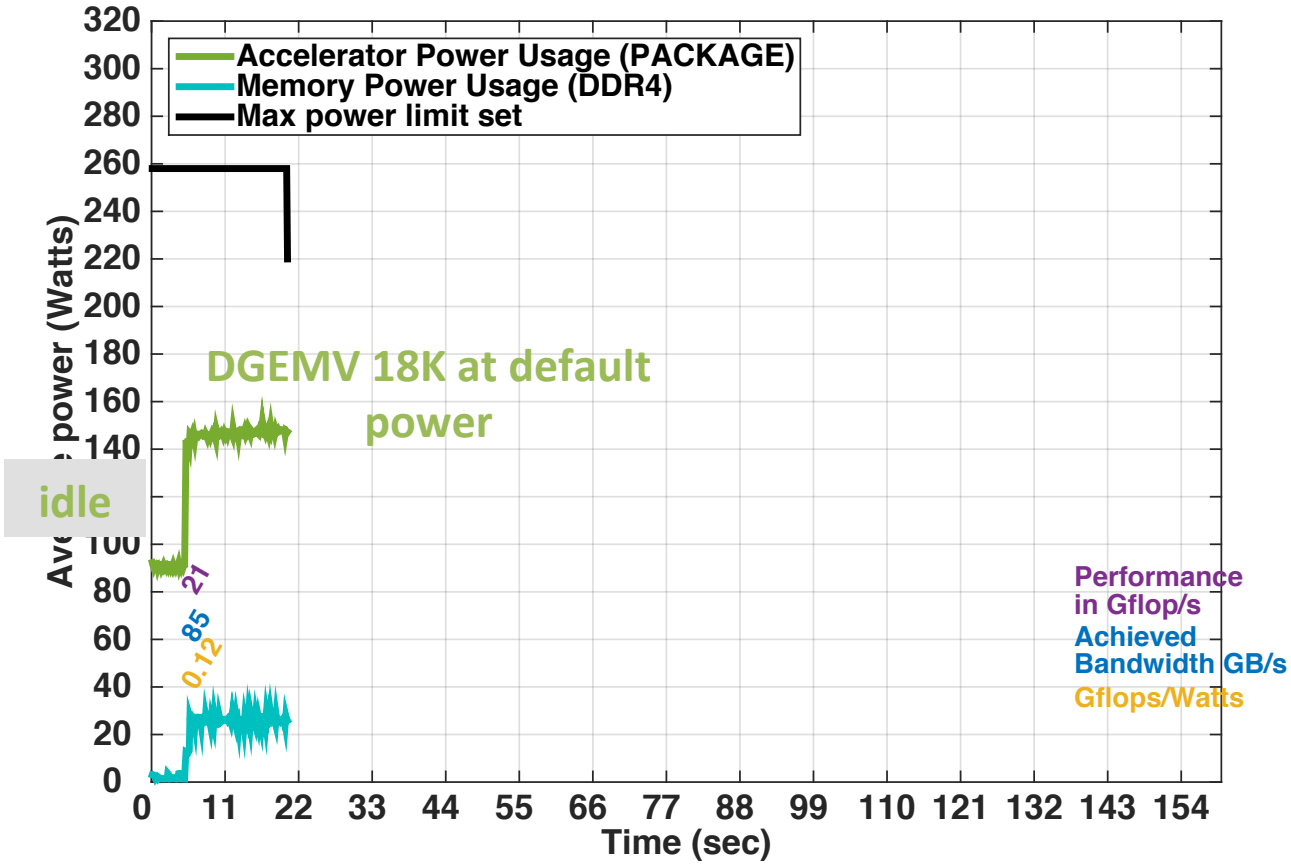
set pwr limit default

DGEMV size 18K

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

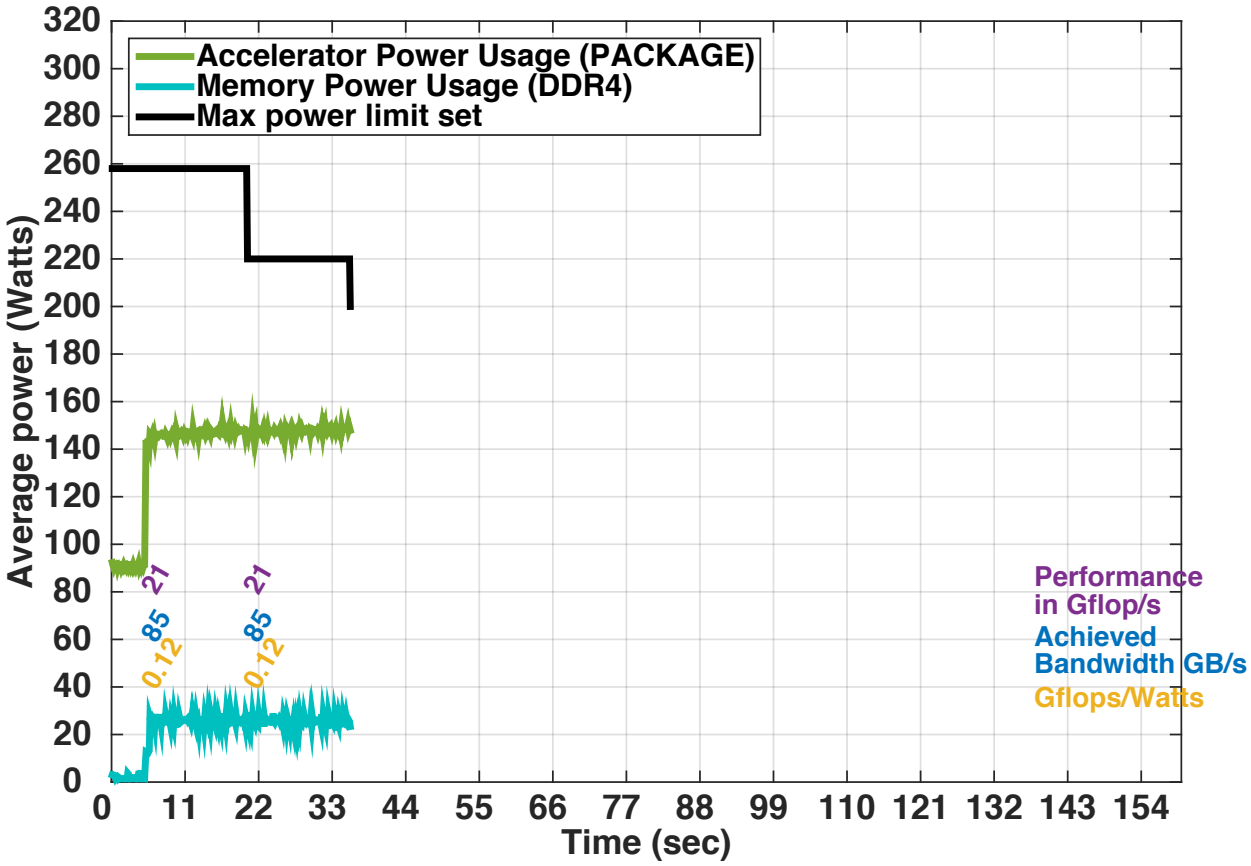


set pwr limit default
DGEMV size 18K

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

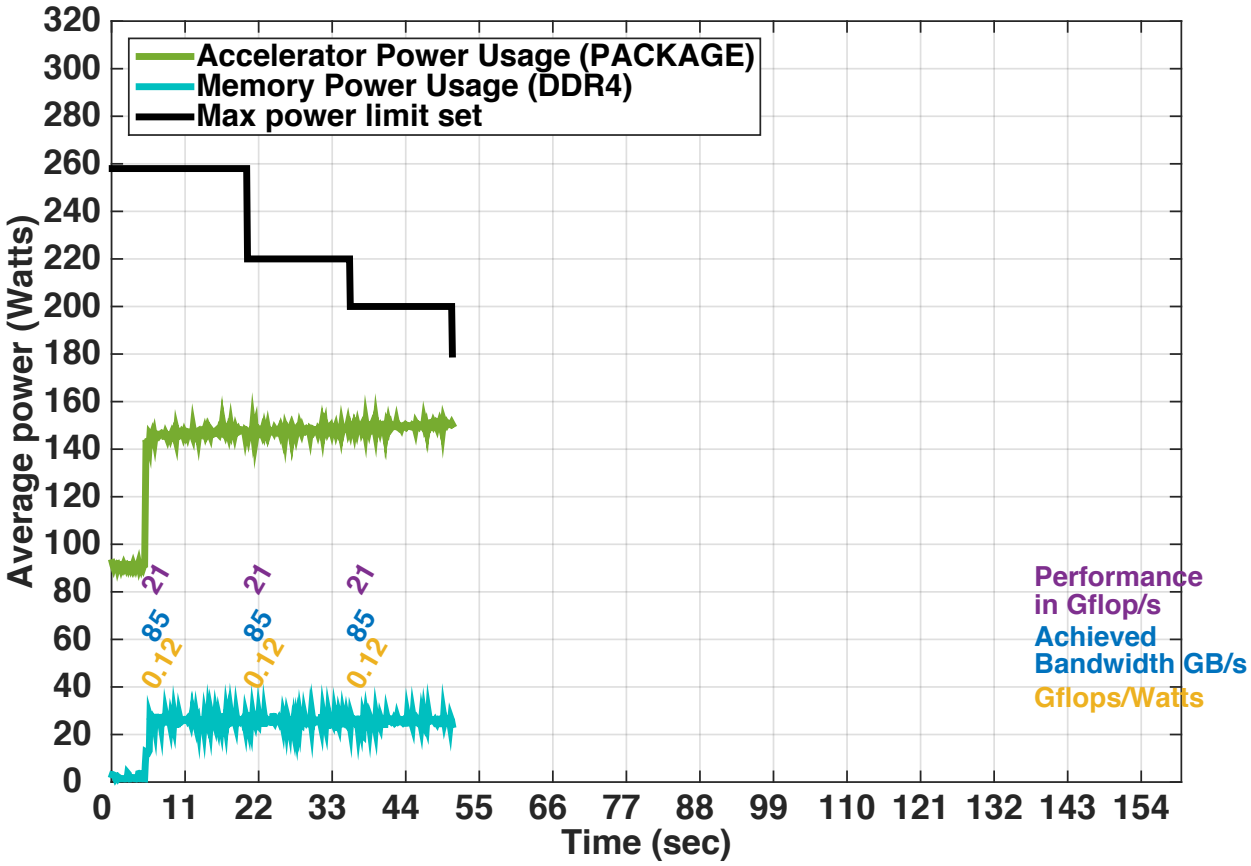


set pwr limit default
DGEMV size 18K
set pwr limit 220W
DGEMV size 18K

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

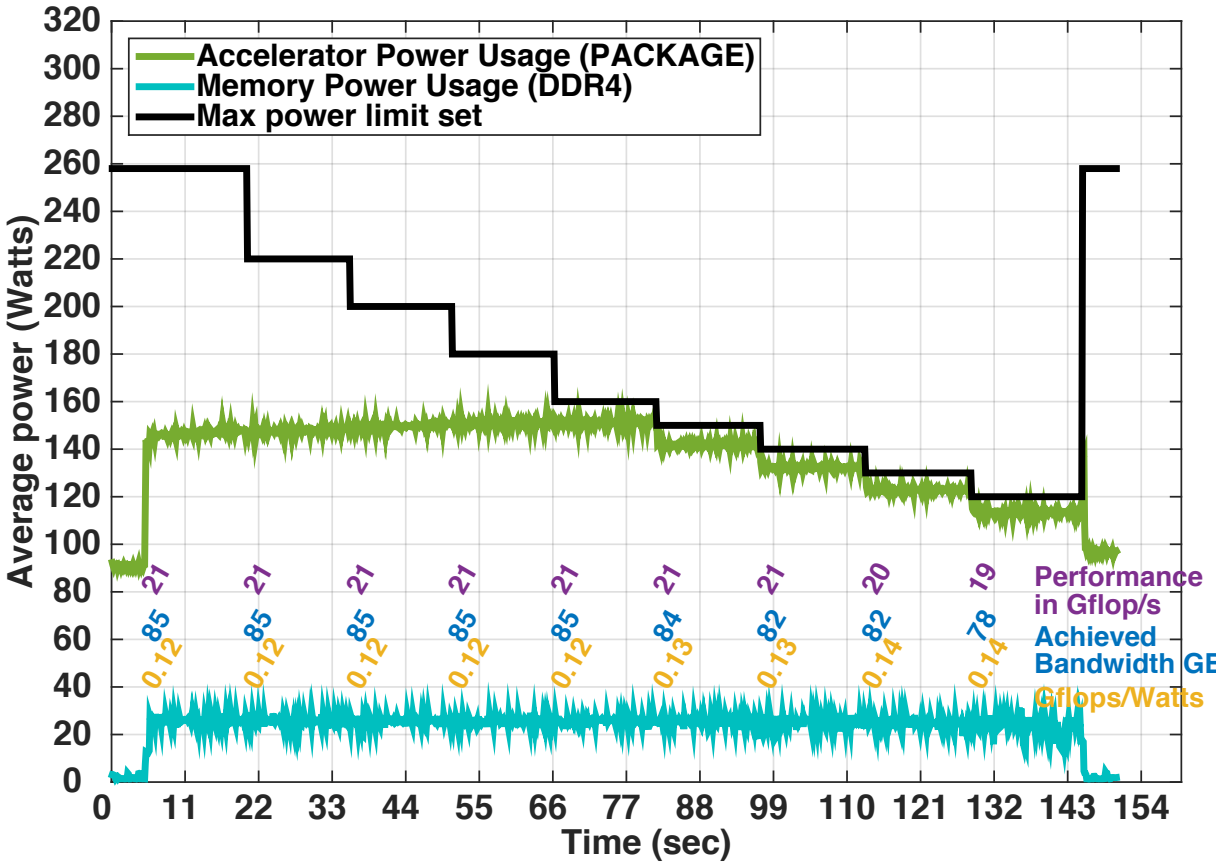


set pwr limit default
DGEMV size 18K
set pwr limit 220W
DGEMV size 18K
set pwr limit 200W
DGEMV size 18K

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

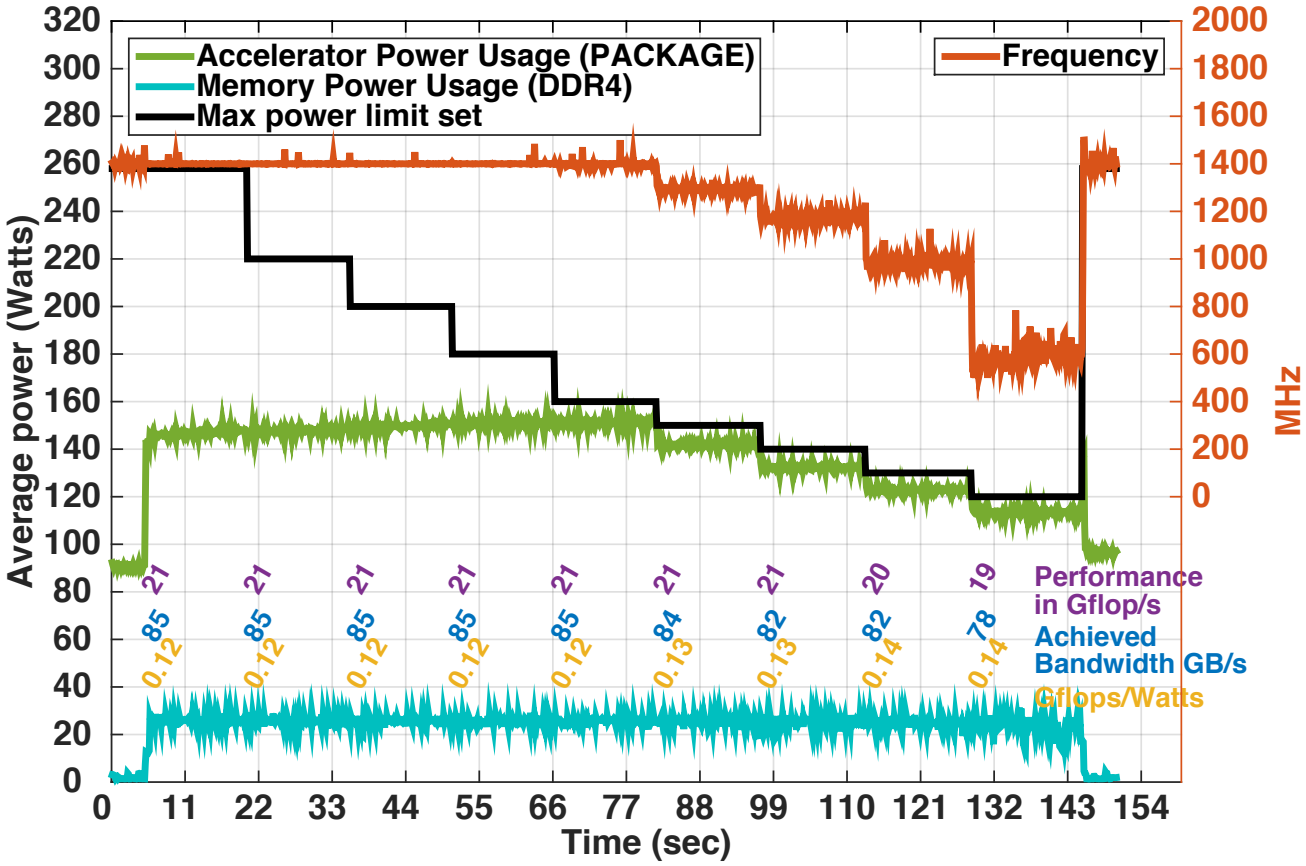


set pwr limit default
DGEMV size 18K
set pwr limit 220W
DGEMV size 18K
set pwr limit 200W
DGEMV size 18K
set pwr limit 180W
DGEMV size 18K
set pwr limit 160W
DGEMV size 18K
set pwr limit 150W
....

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

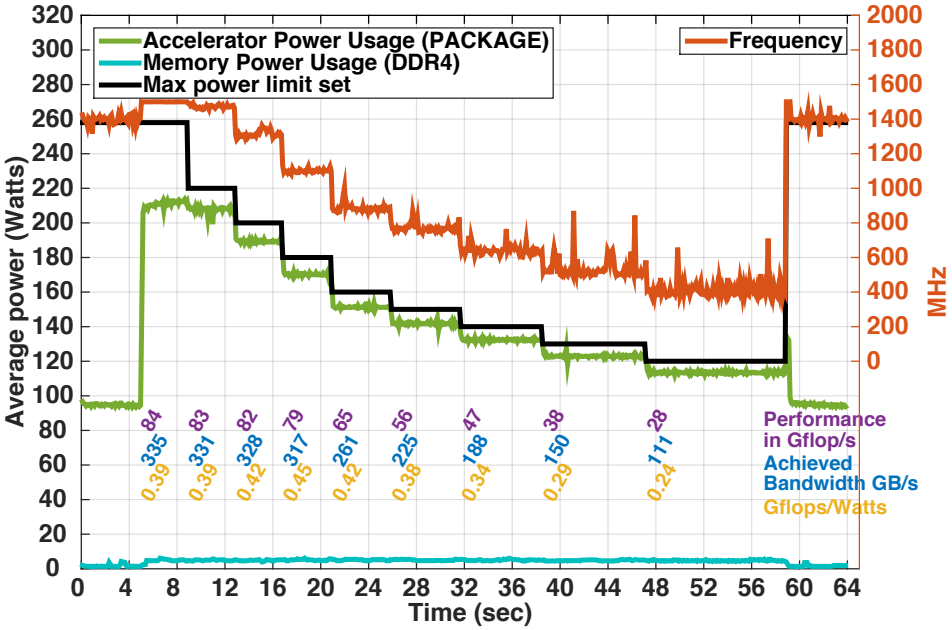


- **Frequency is not affected until the cap kick-on**

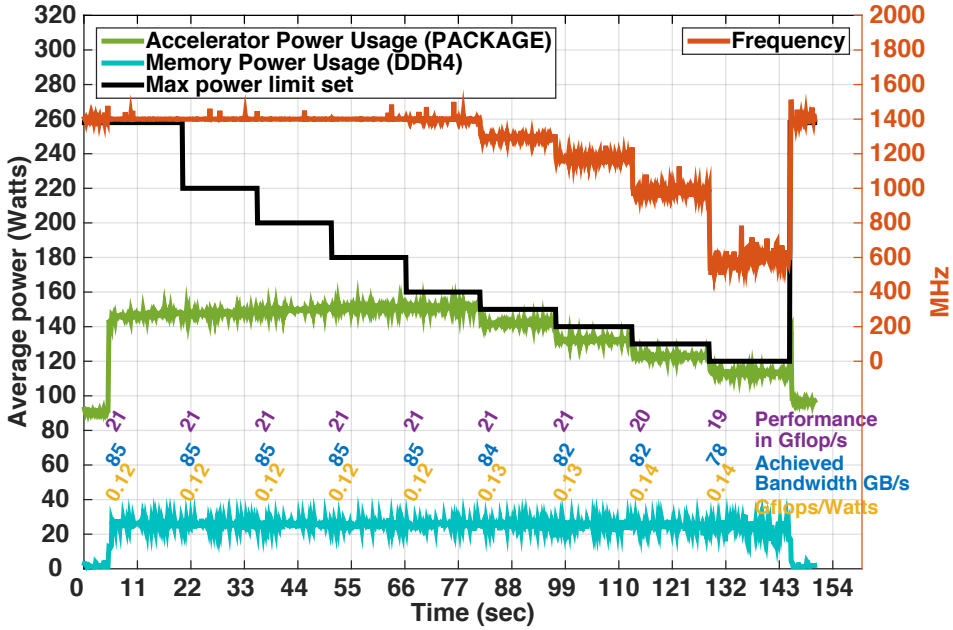
DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 2 BLAS DGEMV on KNL MCDRAM/DDR4

MCDRAM



DDR4



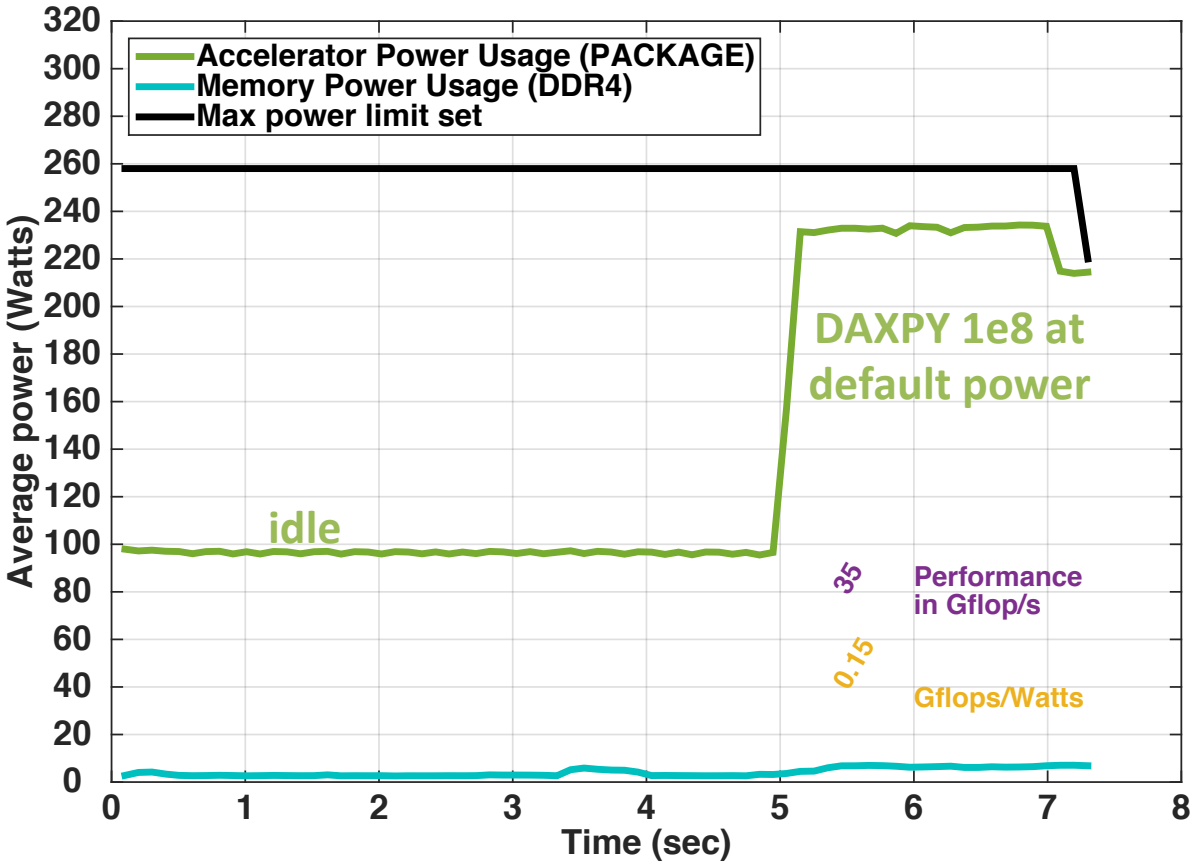
Lesson for DGEMV type of operations (memory bound):

- **For MCDRAM, capping at 190 Watts results in power reduction without any loss if performance.**
- **For DDR4, capping at 130 Watts improves power efficiency by ~20%.**
- **Overall capping 40 Watts below the observed power at default setting provide about 20% reduction without any loss in time to solution.**

DGEMV is run repeatedly for a fixed matrix size of 18K per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

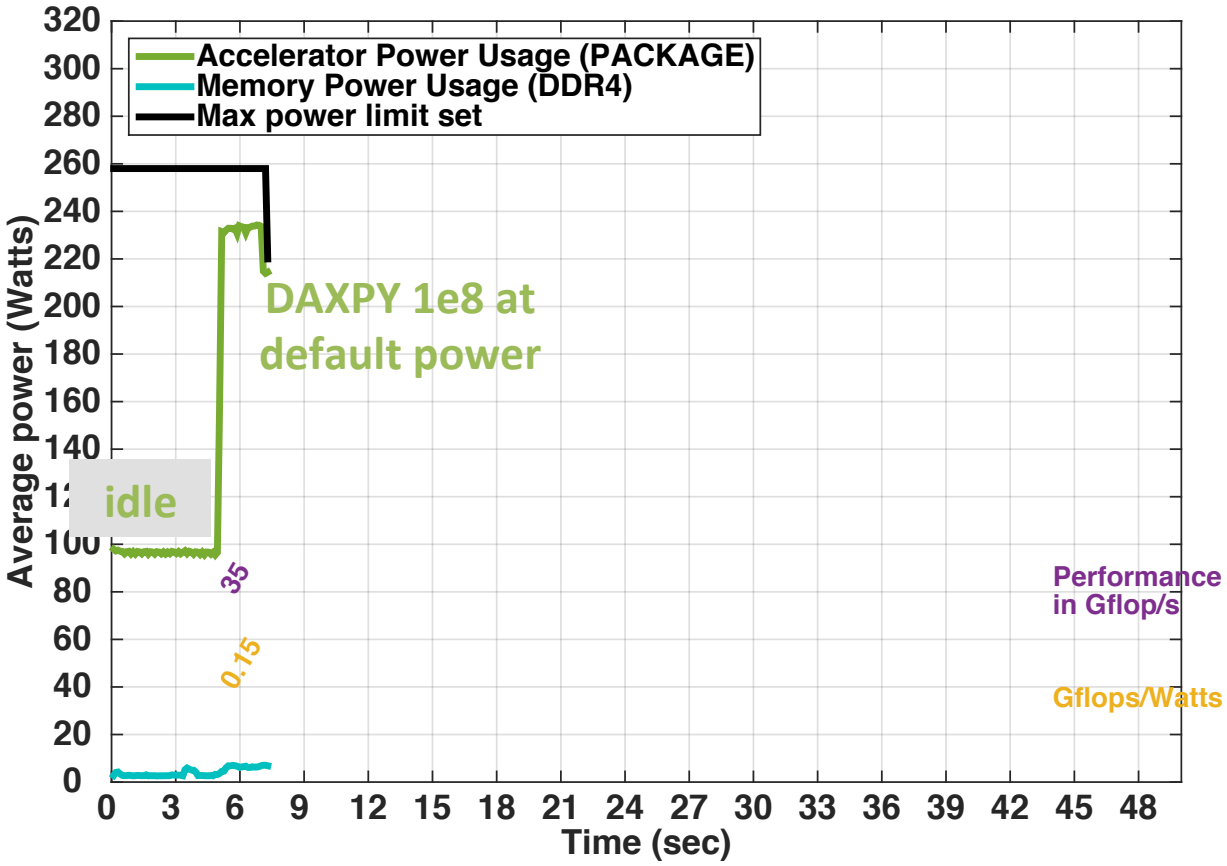


set pwr limit default
DAXPY size 1e8

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s



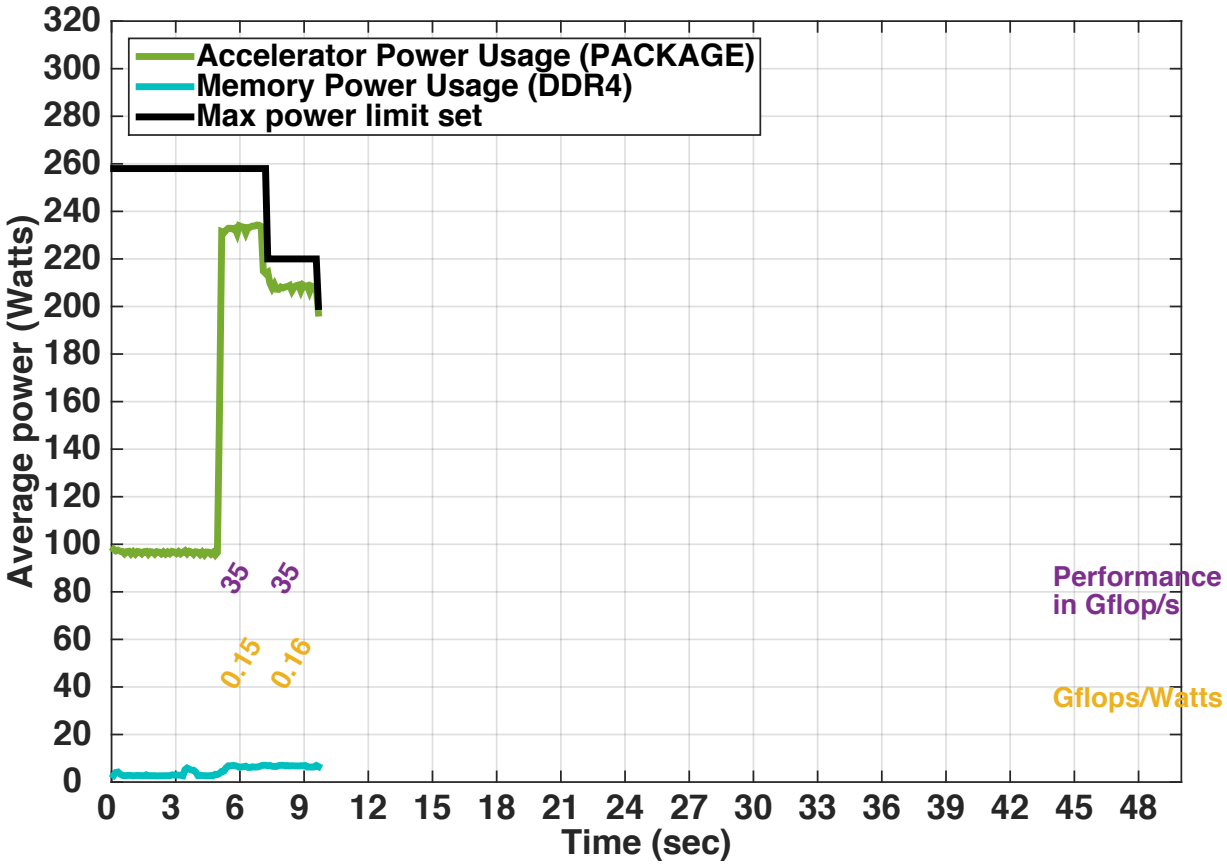
set pwr limit default

DAXPY size 1e8

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

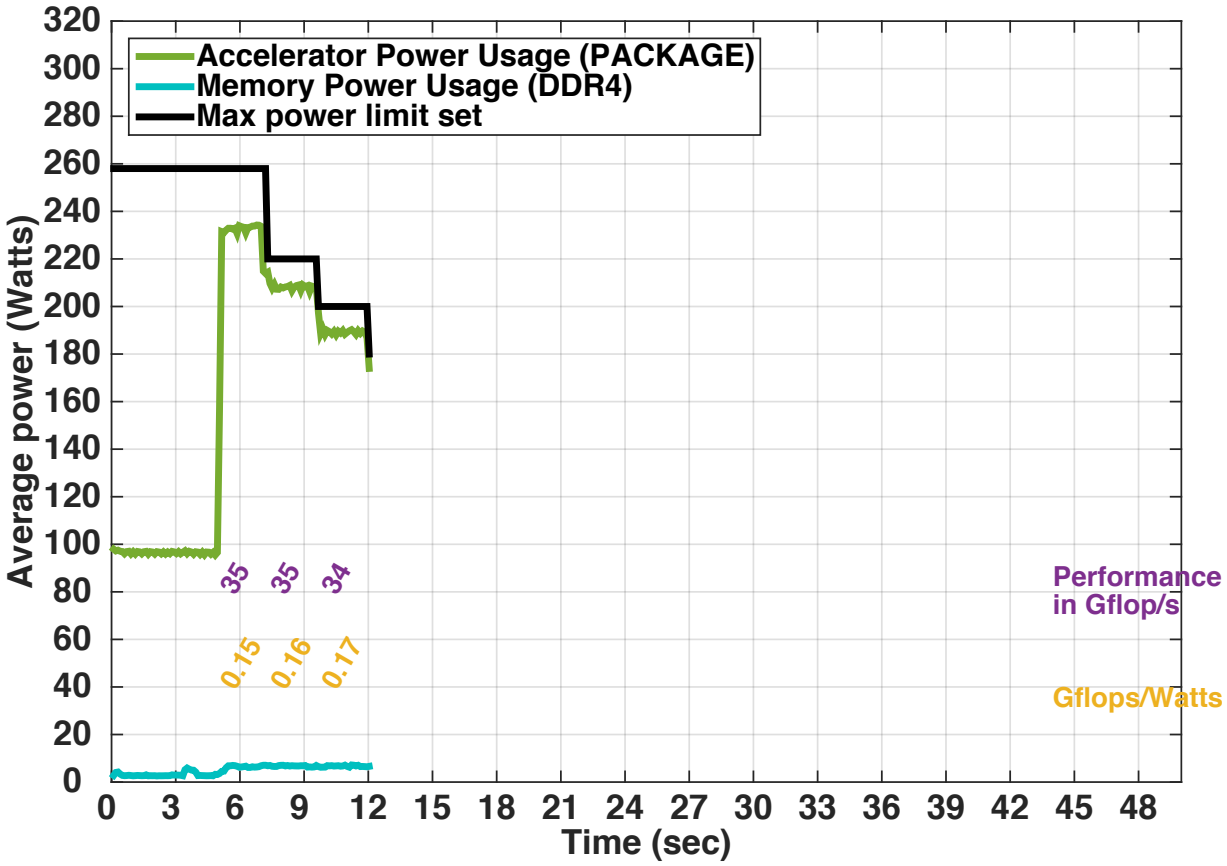


set pwr limit default
DAXPY size 1e8
set pwr limit 220W
DAXPY size 1e8

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

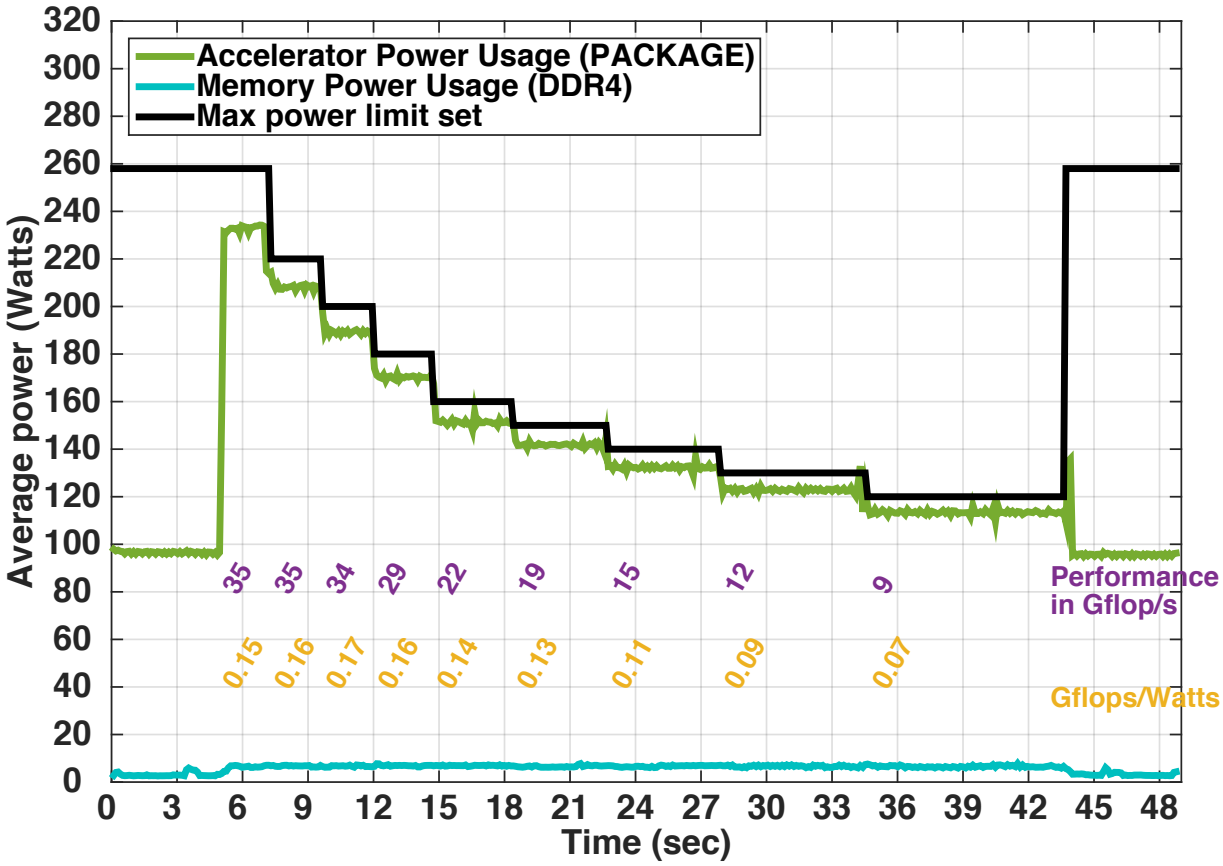


set pwr limit default
DAXPY size 1e8
set pwr limit 220W
DAXPY size 1e8
set pwr limit 200W
DAXPY size 1e8

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

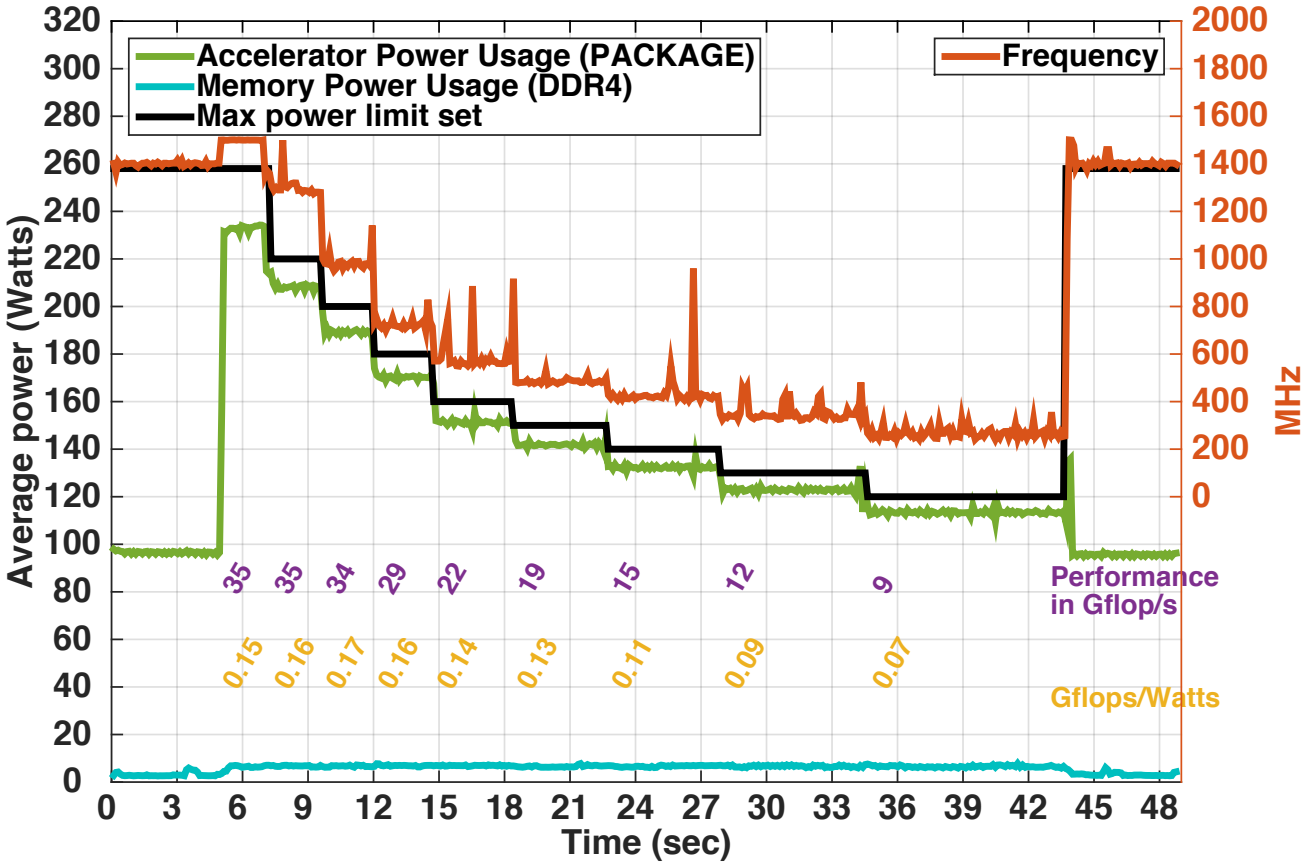


set pwr limit default
DAXPY size 1e8
set pwr limit 220W
DAXPY size 1e8
set pwr limit 200W
DAXPY size 1e8
set pwr limit 180W
DAXPY size 1e8
set pwr limit 160W
DAXPY size 1e8
set pwr limit 150W
DAXPY size 1e8
....

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using MCDRAM

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

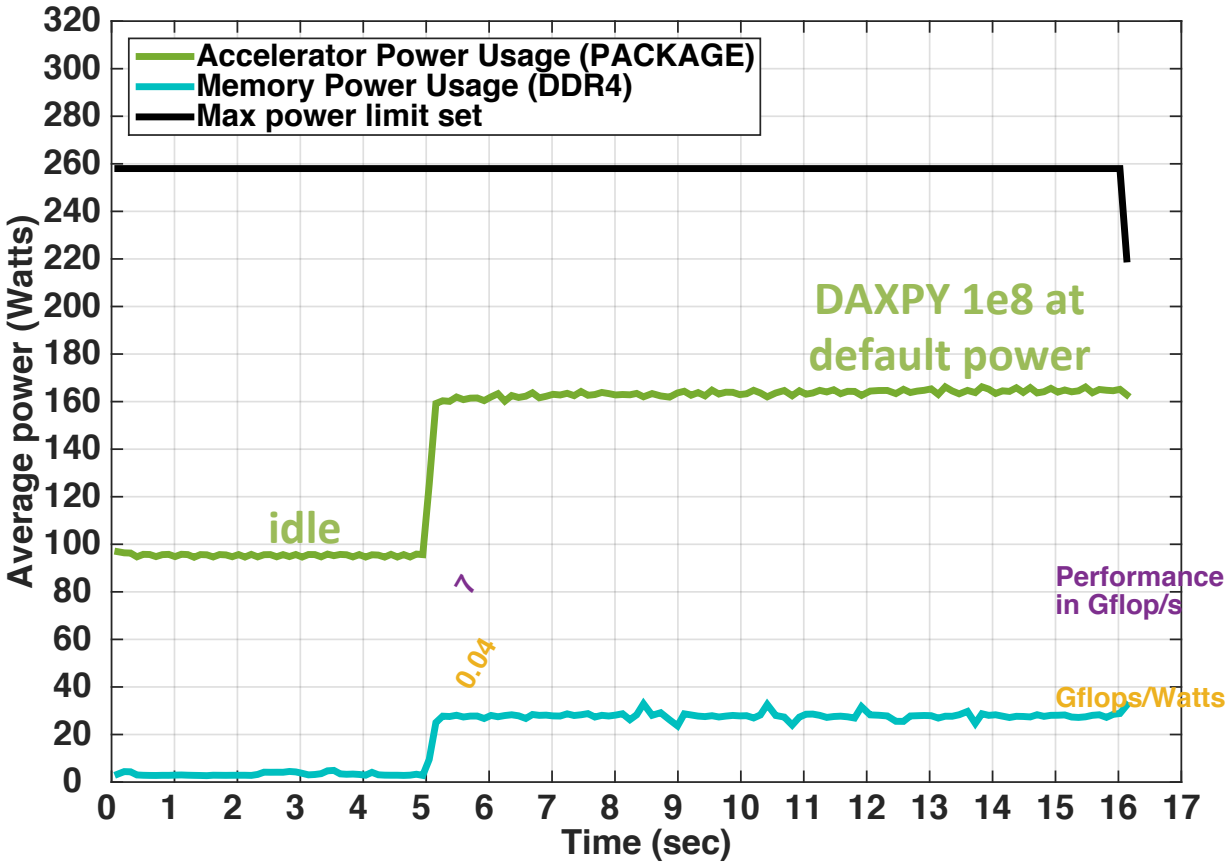


- **Frequency is not affected until the cap kick-on**

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

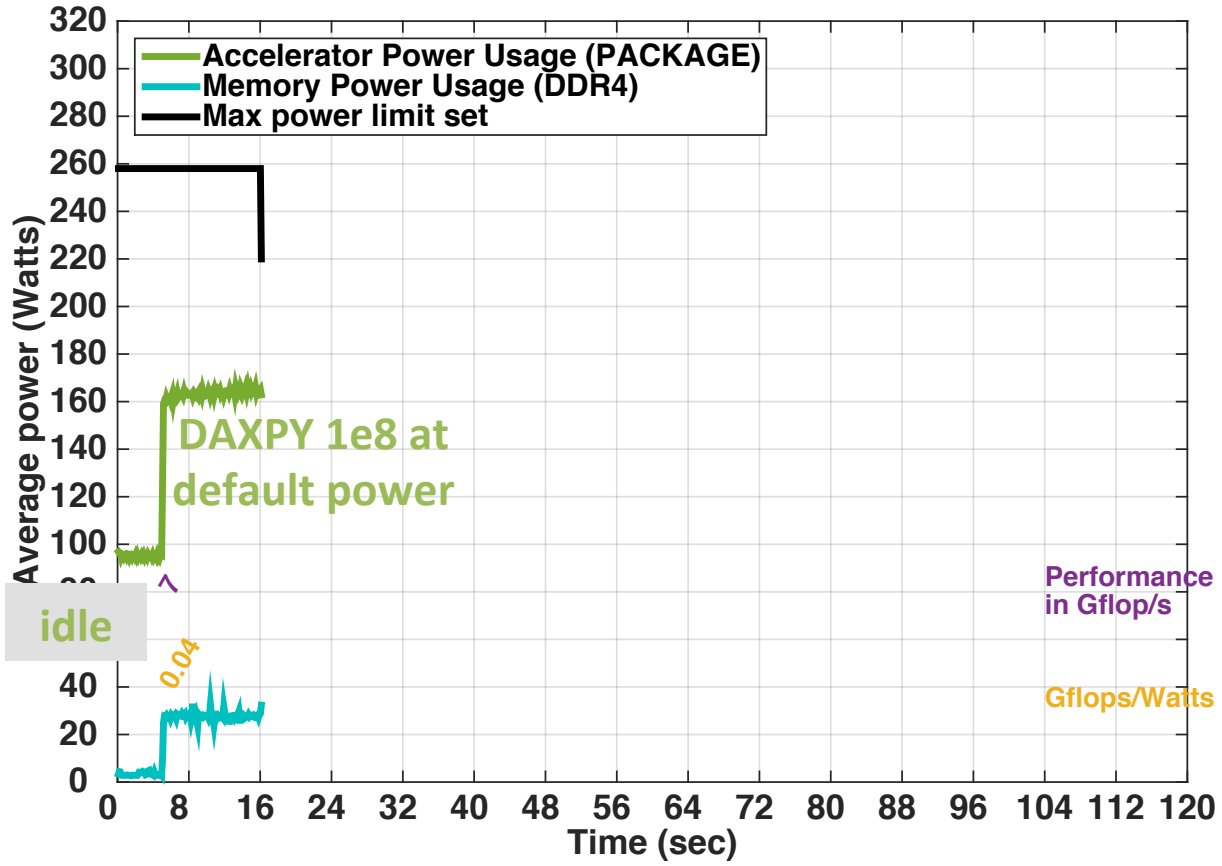


set pwr limit default
DAXPY size 1e8

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

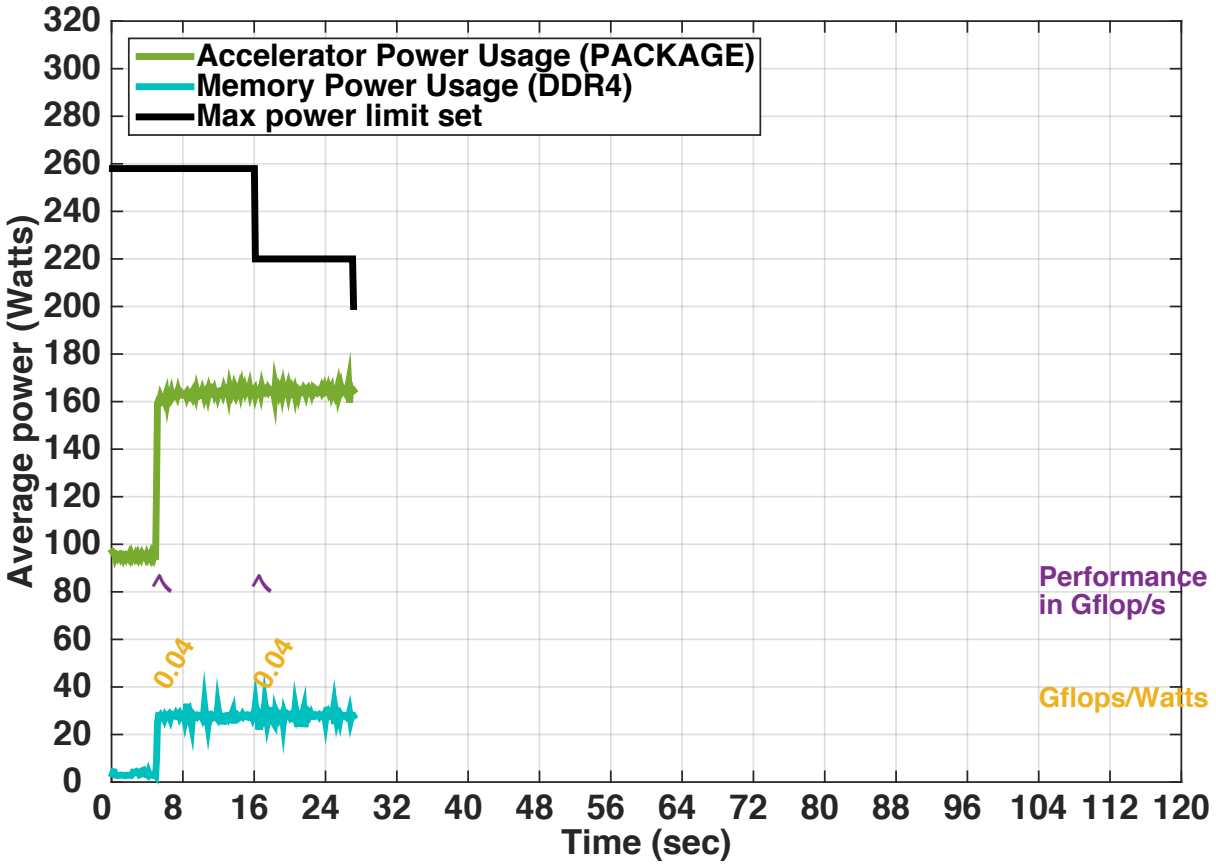


set pwr limit default
DAXPY size 1e8

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

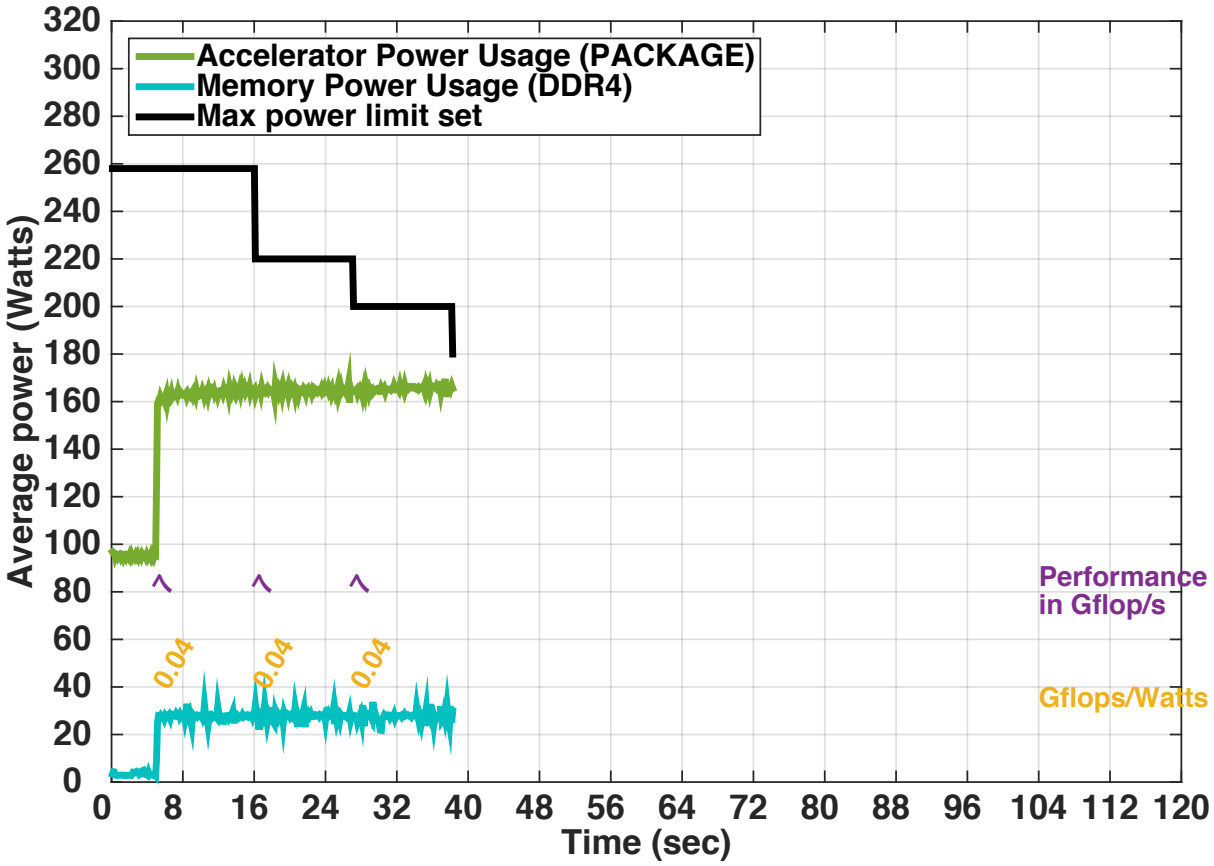


set pwr limit default
DAXPY size 1e8
set pwr limit 220W
DAXPY size 1e8

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

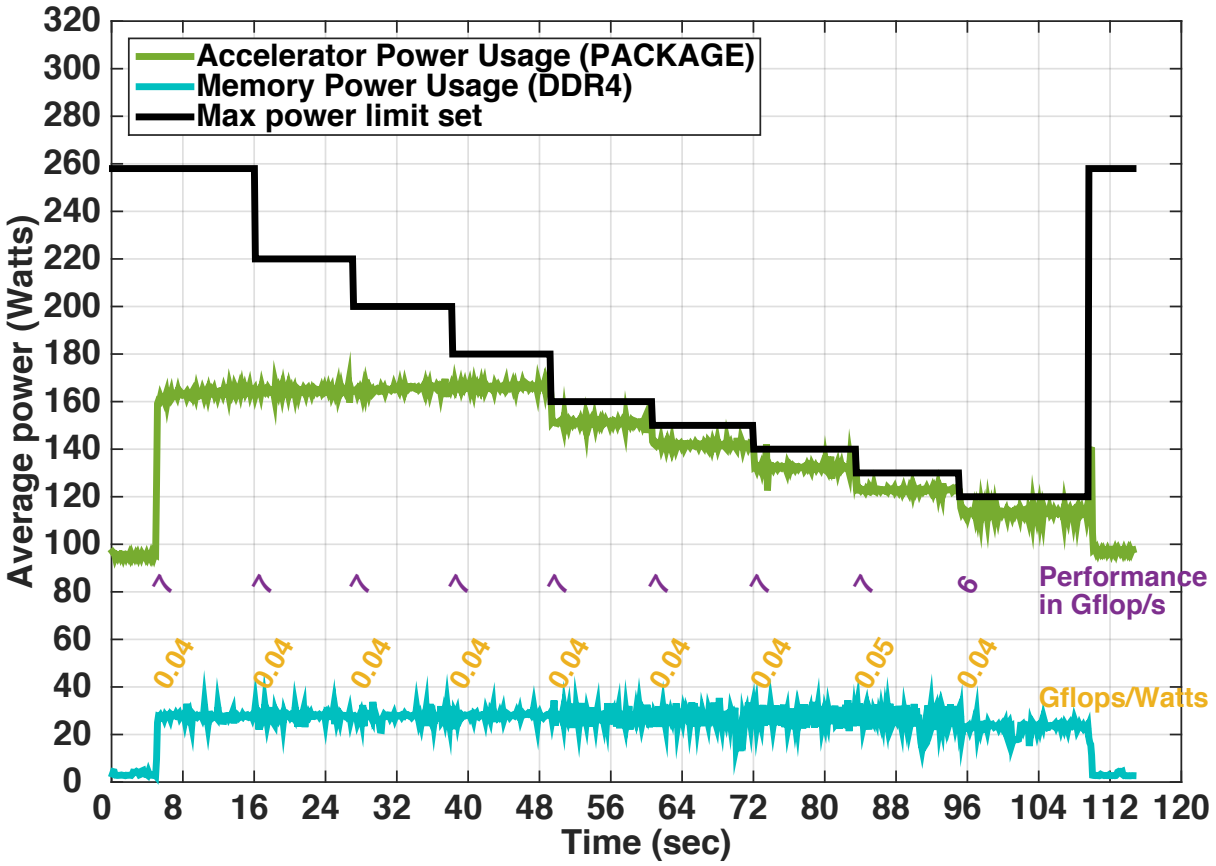


set pwr limit default
DAXPY size 1e8
set pwr limit 220W
DAXPY size 1e8
set pwr limit 200W
DAXPY size 1e8

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

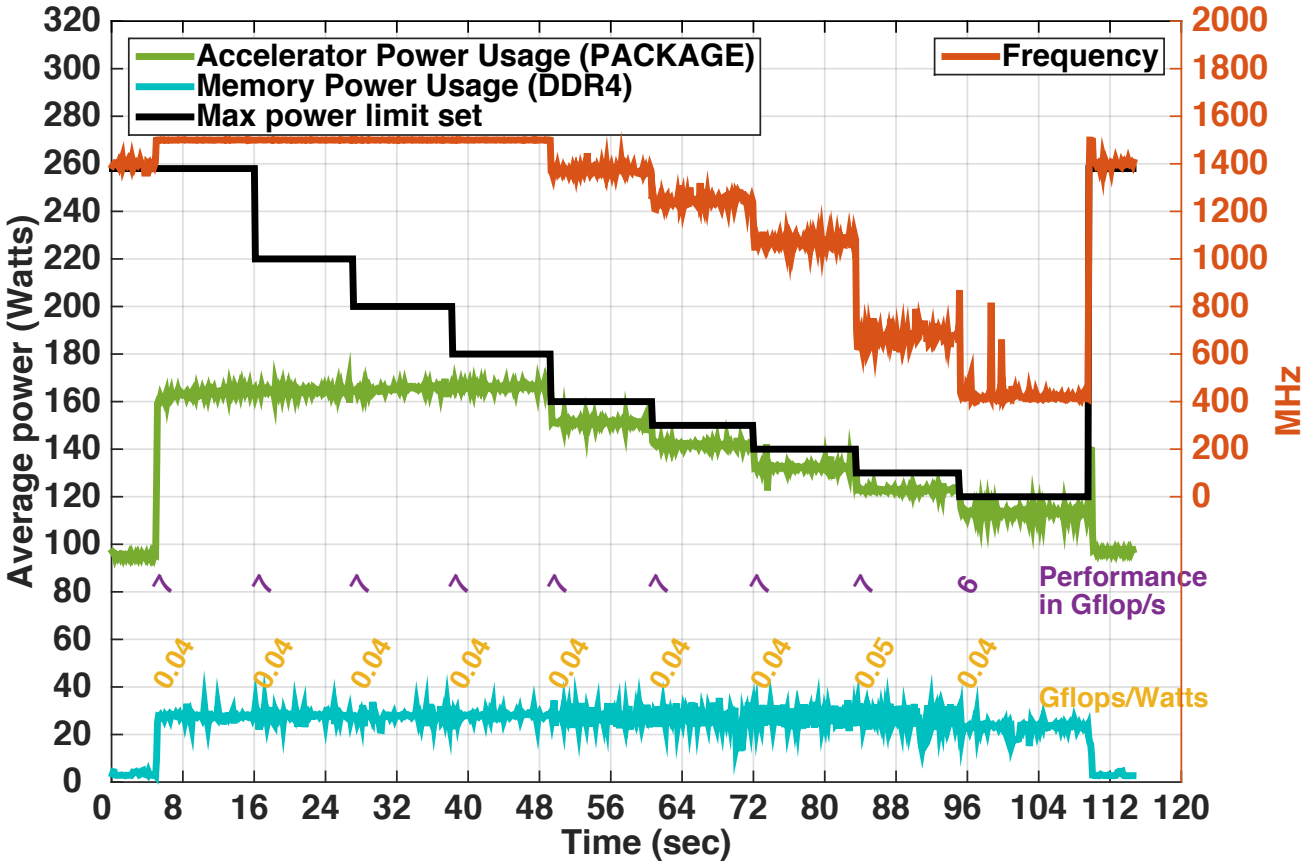


set pwr limit default
DAXPY size 1e8
set pwr limit 220W
DAXPY size 1e8
set pwr limit 200W
DAXPY size 1e8
set pwr limit 180W
DAXPY size 1e8
set pwr limit 160W
DAXPY size 1e8
set pwr limit 150W
DAXPY size 1e8
....

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL using DDR4

68 cores **KNL**, Peak DP = 2662 Gflop/s Bandwidth **MCDRAM ~425 GB/s** DDR4 ~90 GB/s

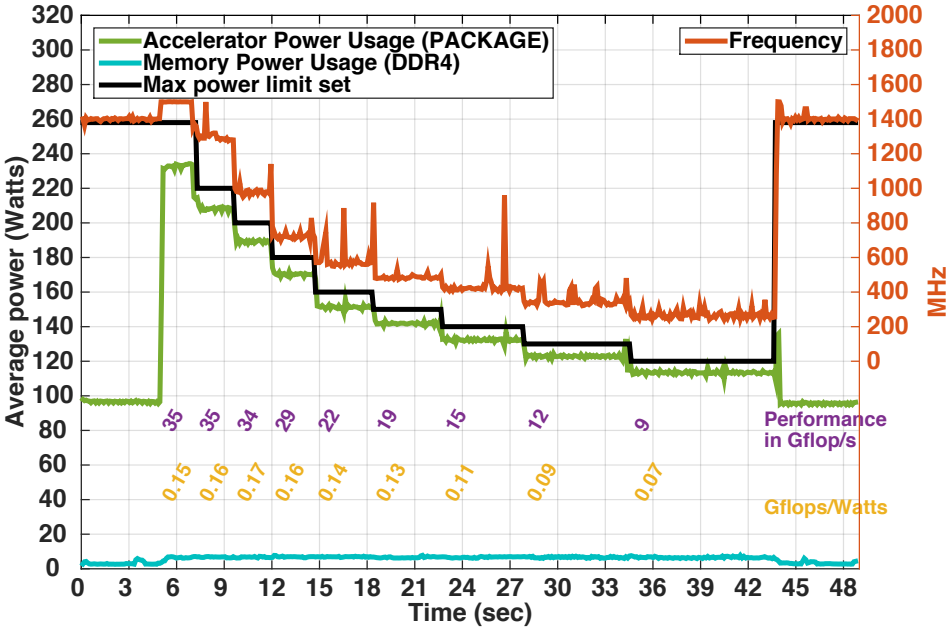


- **Frequency is not affected until the cap kick-on**

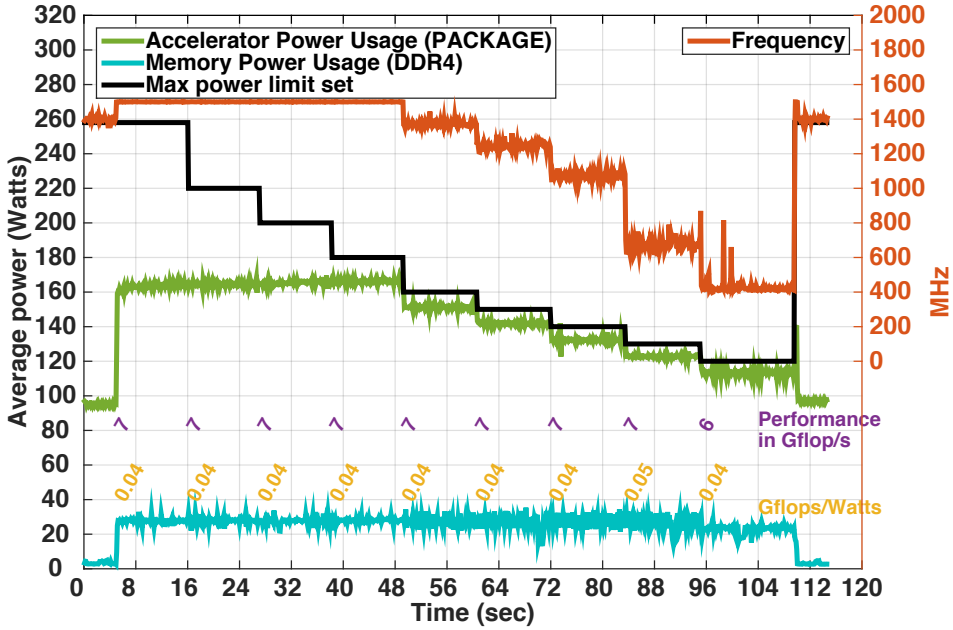
DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Level 1 BLAS DAXPY on KNL MCDRAM/DDR4

MCDRAM



DDR4



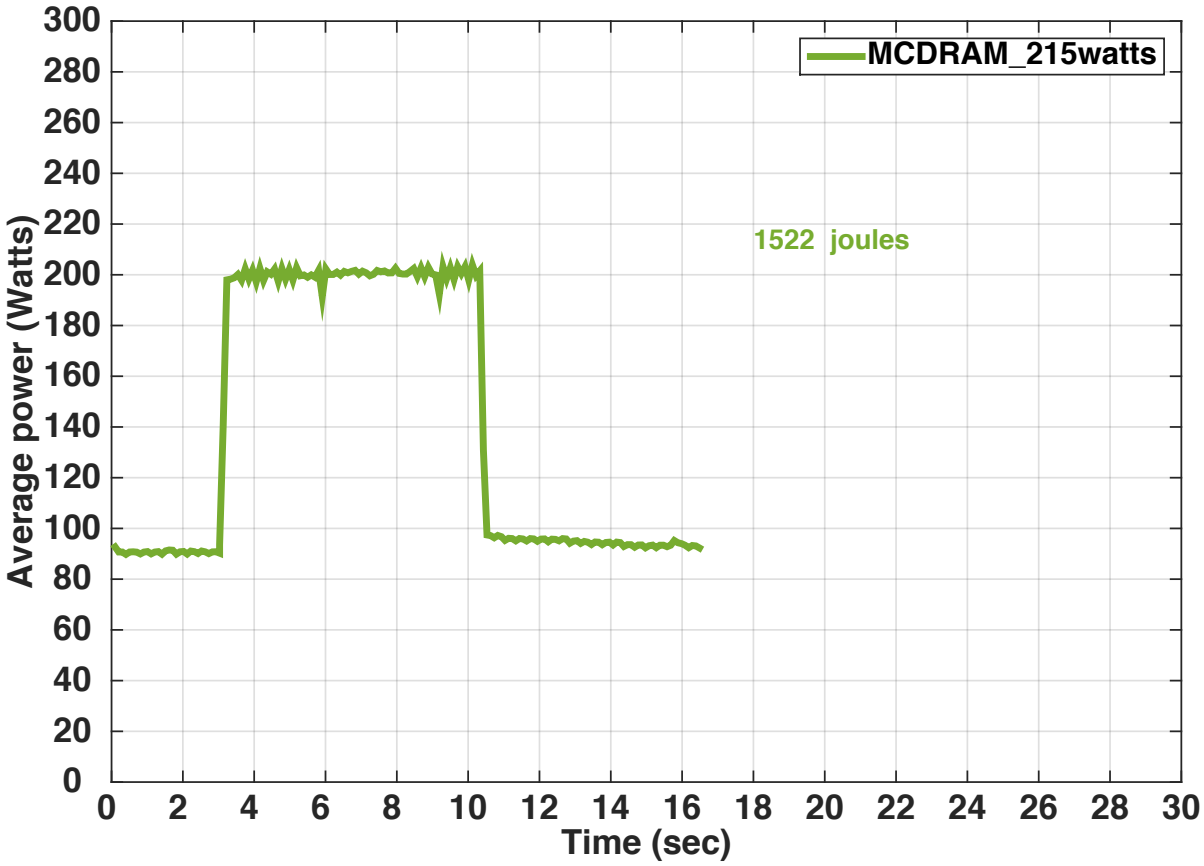
Lesson for DAXPY type of operations (memory bound):

- **For MCDRAM, capping at 170 Watts results in power reduction without any loss in performance.**
- **For DDR4, capping at 130 Watts improves power efficiency by ~20%.**
- **Overall capping 40 Watts below the observed power at default setting provide about 20% reduction without any loss in time to solution.**

DAXPY is run repeatedly for a fixed matrix size of 1E8 per step and at each step a new power limit is set in decreasing fashion starting from default, 220 Watts, 200 Watts down till 120 Watts by steps of 10/20.

Power-awareness in applications

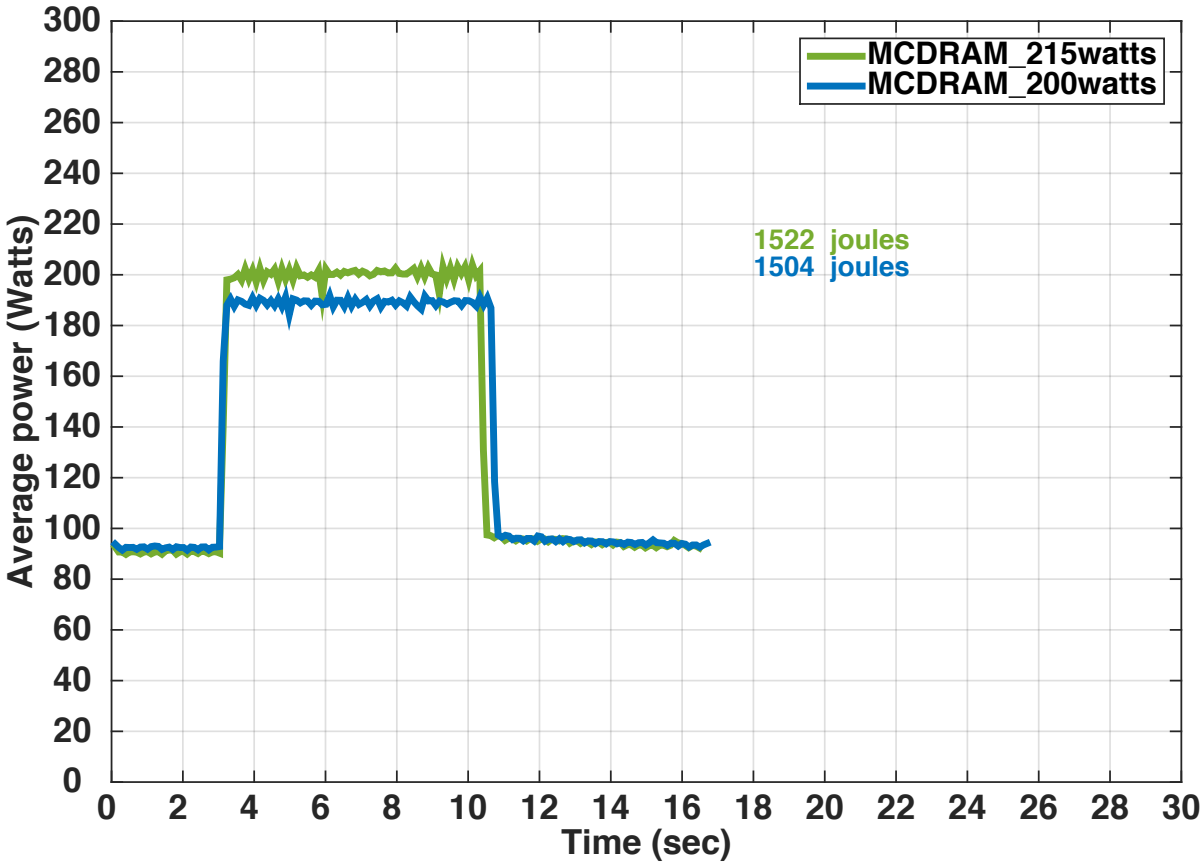
HPCG benchmark on grid of size 192^3: MCDRAM



At TDP basic power limit 215

Power-awareness in applications

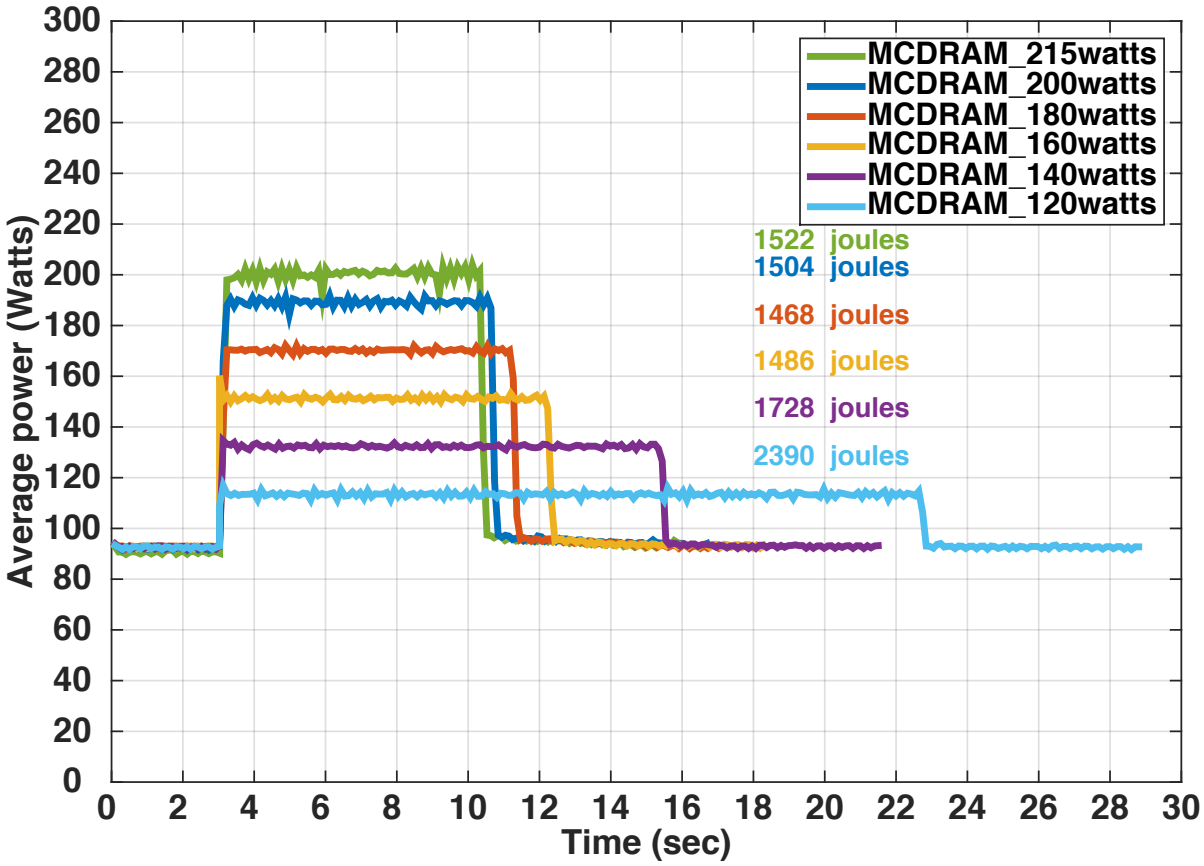
HPCG benchmark on grid of size 192^3: MCDRAM



- Decreasing the power limit to 200 do not results in any loss in performance while we observe a reducing of power consumption

Power-awareness in applications

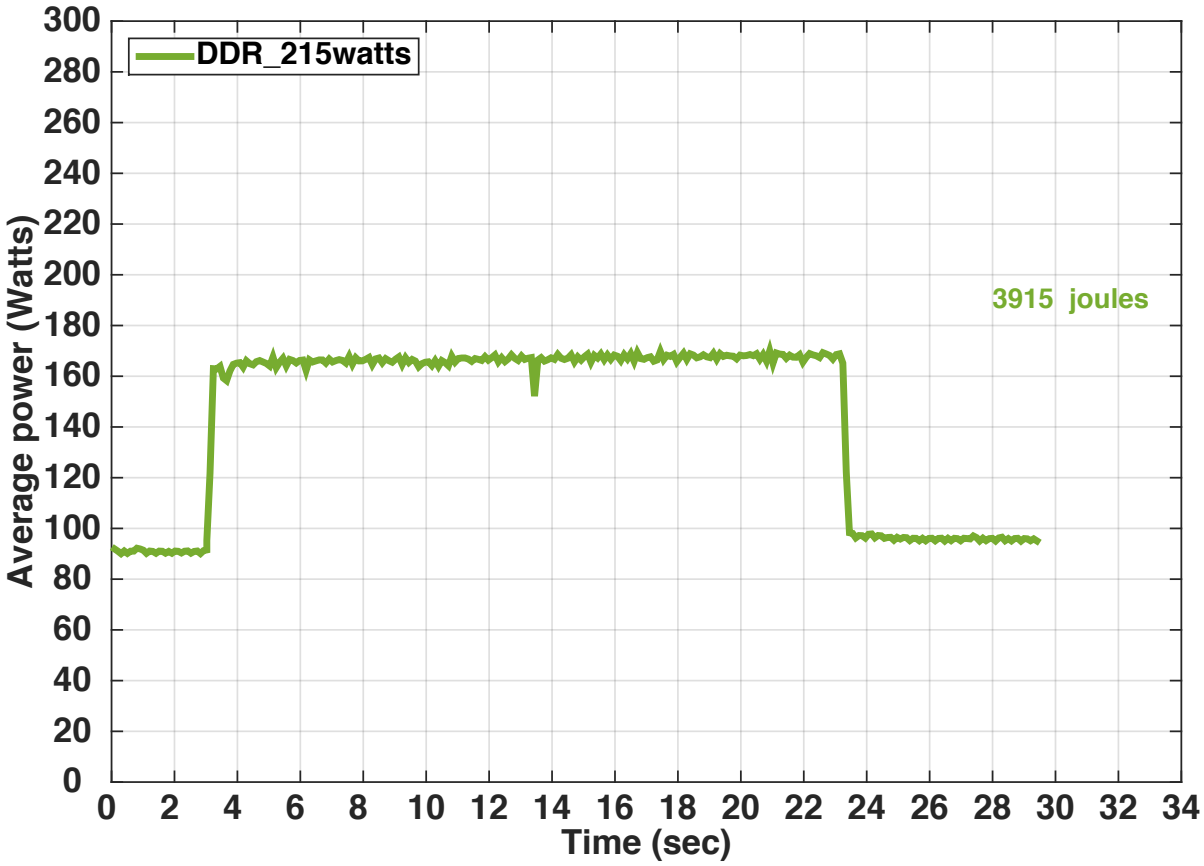
HPCG benchmark on grid of size 192^3: MCDRAM



- Decreasing the power limit to 200 do not results in any loss in performance while we observe a reducing of power consumption.
- Decreasing the power limit down by 40 Watts (Pwr limit at 160) will keep provide power reduction, large energy saving and without any practical loss in performance, less than 10% loss in time to solution.

Power-awareness in applications

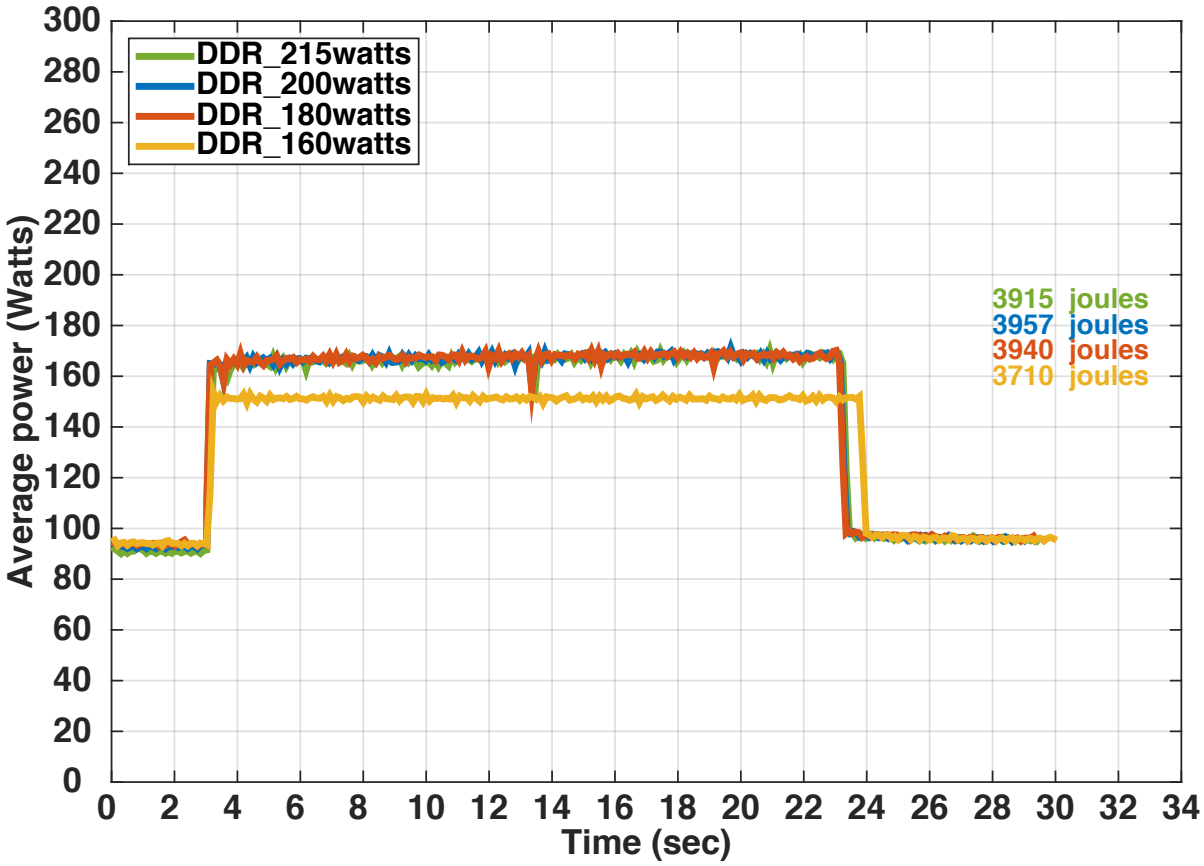
HPCG benchmark on grid of size 192^3: DDR4



- At TDP basic power limit 215

Power-awareness in applications

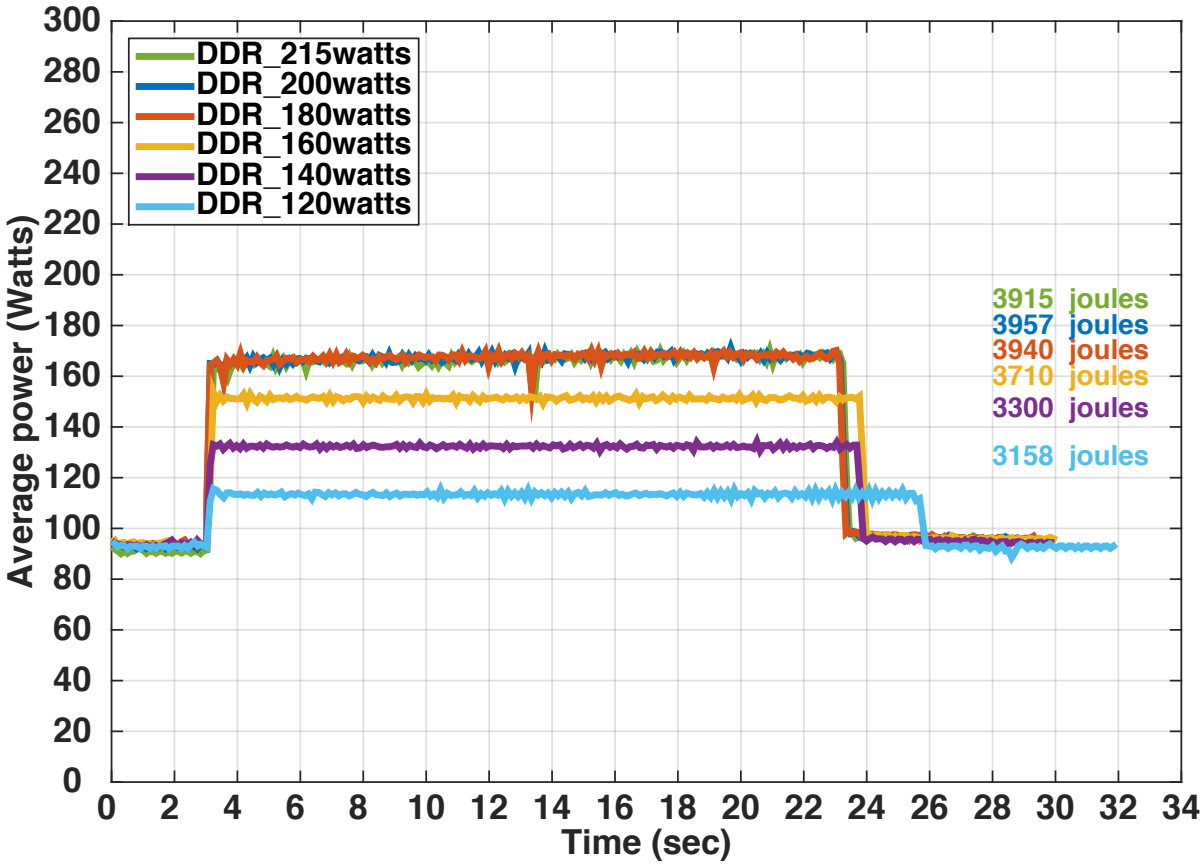
HPCG benchmark on grid of size 192^3: DDR4



- Decreasing the power limit to 160 do not results in any loss in performance while we observe a reducing of power consumption

Power-awareness in applications

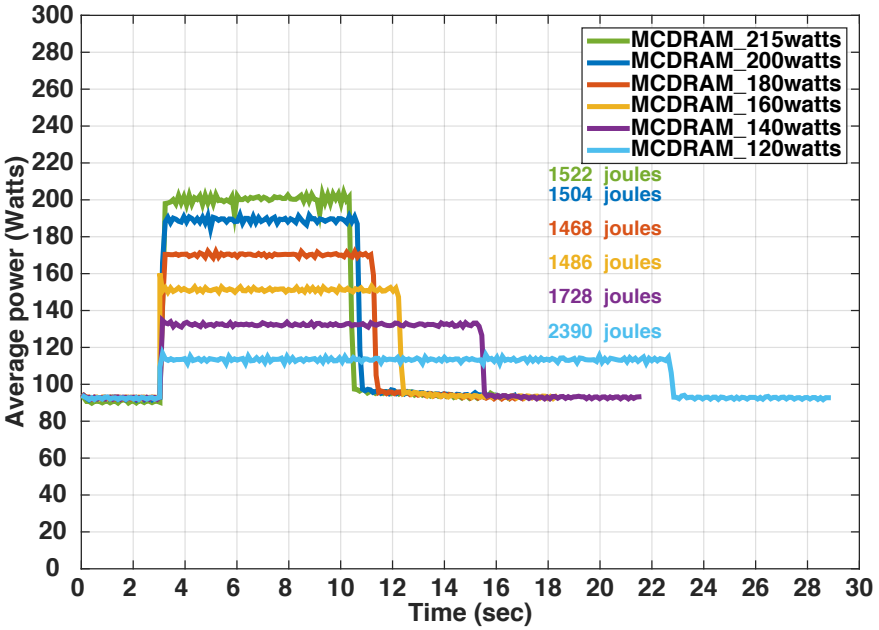
HPCG benchmark on grid of size 192^3: DDR4

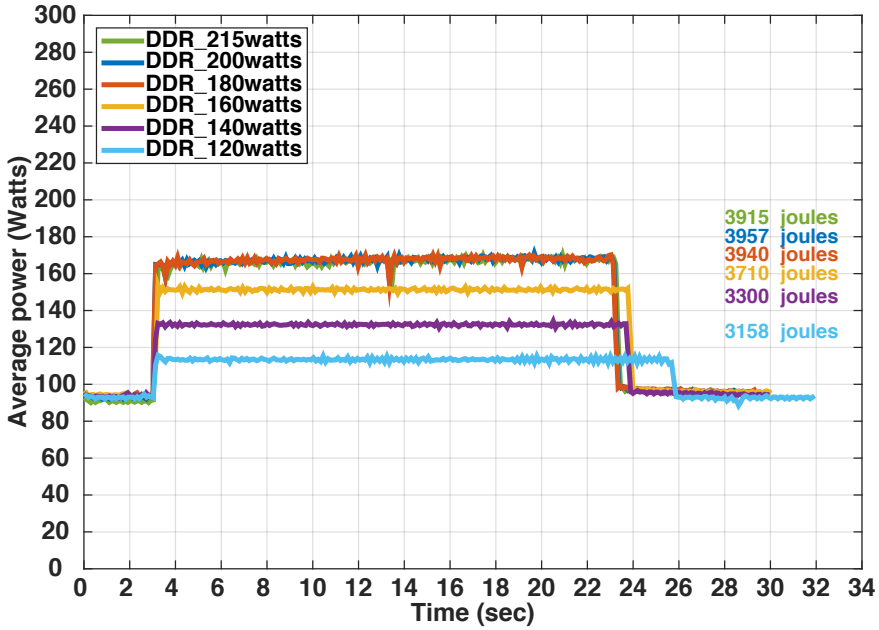


- Decreasing the power limit to 200 do not results in any loss in performance while we observe a reducing of power consumption.
- Decreasing the power limit down by 40 Watts (Pwr limit at 140) will keep providing power reduction, large energy saving and without any practical loss in performance.
- At 120 Watts, less than 10% loss in time to solution while a large energy and power savings is observed

Power-awareness in applications

HPCG benchmark on grid of size 192^3:





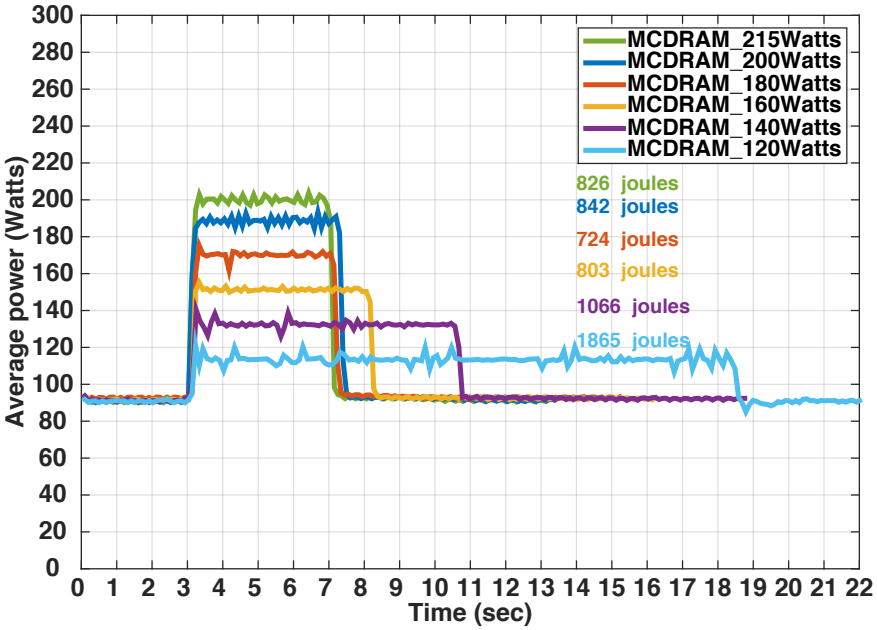
Lesson for HPCG:

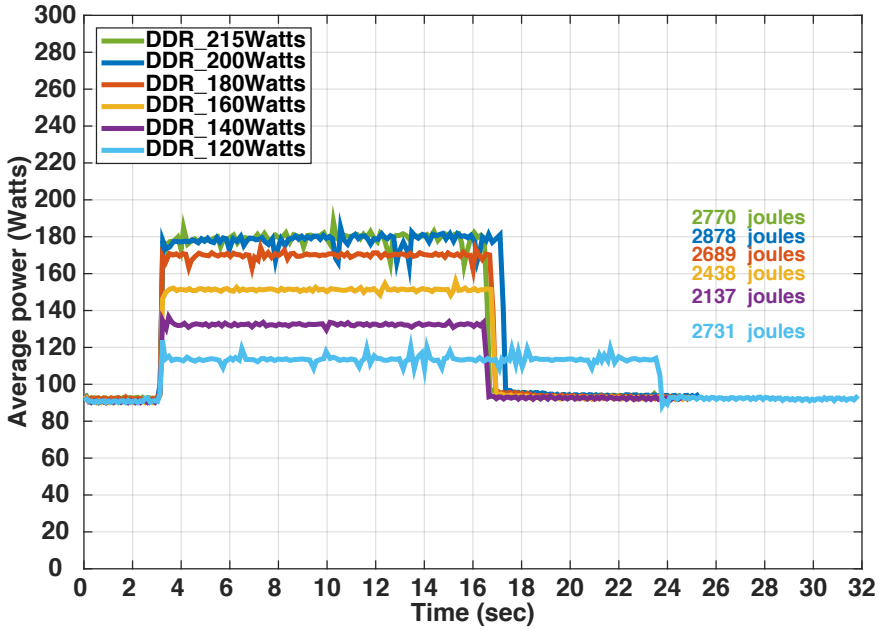
- For MCDRAM, decreasing the power limit by 40 Watts from the observed power consumption at default TDP (200→ 160) provide power reduction, large energy saving and without any practical loss in performance.

- For DDR4, similar behavior observed by decreasing the power limit by 40 Watts from the observed power consumption at default TDP (170 → 130) which provide power reduction, large energy saving and without any practical loss in performance.

Power-awareness in applications

Solving Helmholtz equation with finite difference, Jacobi iteration 12800x12800 grid



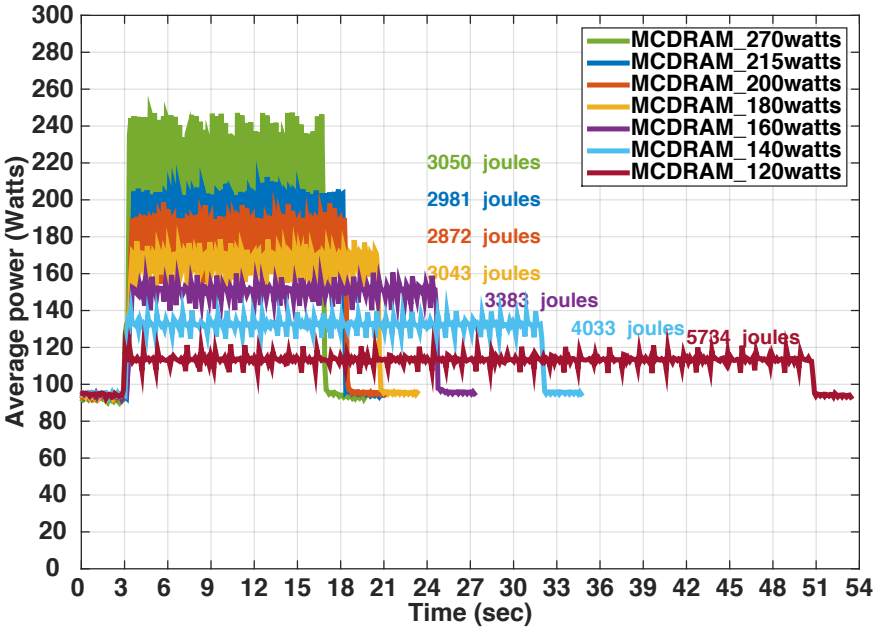


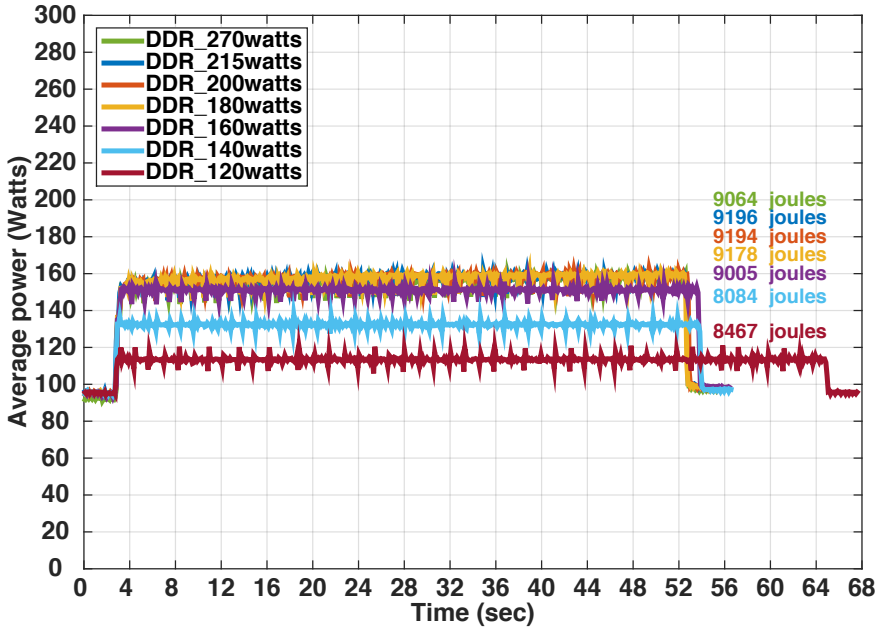
Lesson for Jacobi iteration:

- For MCDRAM, capping at 180-160 Watts Watts improves power efficiency by ~20% without any loss in time to solution.
- For DDR4, capping at 140 Watts improves power efficiency by ~20% without any loss in time to solution.
- Overall 40 Watts below the observed power at default limit is also providing large energy gain while keeping up with the same time to solution which is similar to the behavior observed in DGEMV and DAXPY.

Power-awareness in applications

Lattice-Boltzmann simulation of CFD (from SPEC 2006 benchmark)



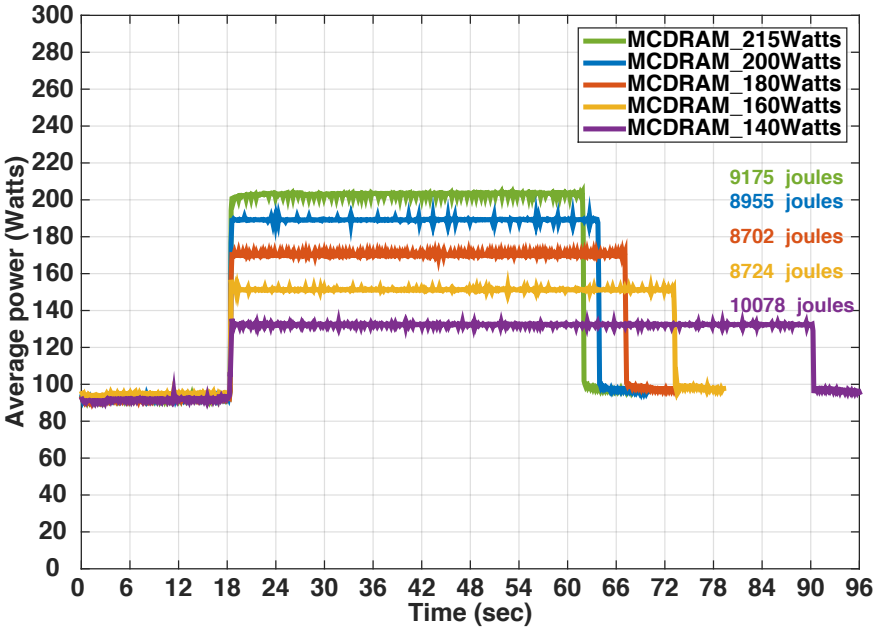


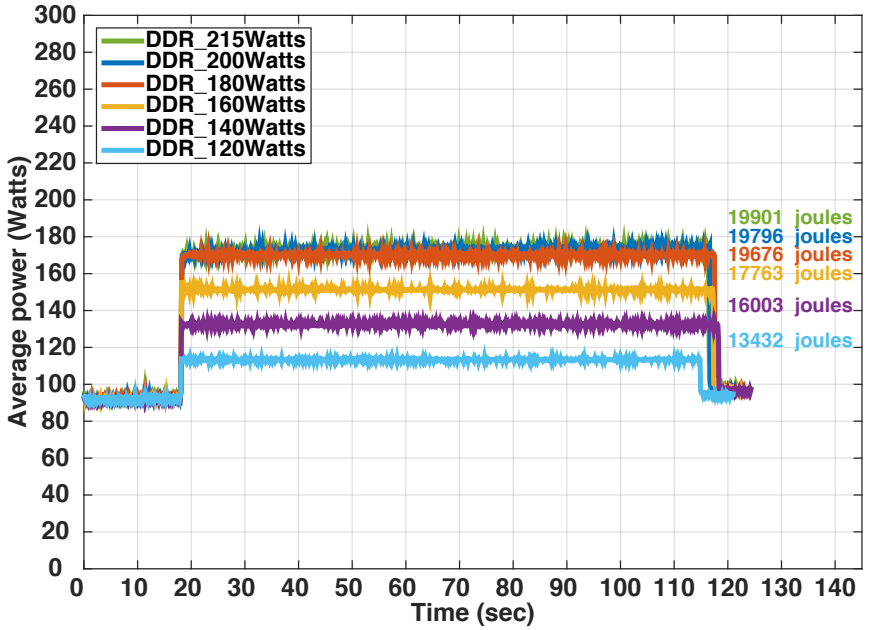
Lesson for Lattice Boltzmann:

- For MCDRAM, capping at 180 results in power reduction with about 10% loss of time to solution.
- For DDR4, capping at 120-140 Watts improves power efficiency by ~20%.

Power-awareness in applications

Monte Carlo neutron transport application (from XSbench)





Lesson for Monte Carlo:

- For MCDRAM, capping at 160 results in power reduction, energy savings with about 10% loss of time to solution.
- For DDR4, capping at 120 Watts improves power efficiency by ~30%.

Power-aware computing: Mixed-precision algorithms

- Savings by Algorithmic advancement, such as mixed precision algorithms.

→ **Objective:** Algorithmic techniques to reduce power consumption by decreasing the execution time → **Energy Savings** !!!

Power-aware computing: Mixed-precision algorithms

- Iterative refinement for dense systems, $Ax = b$, can work this way.

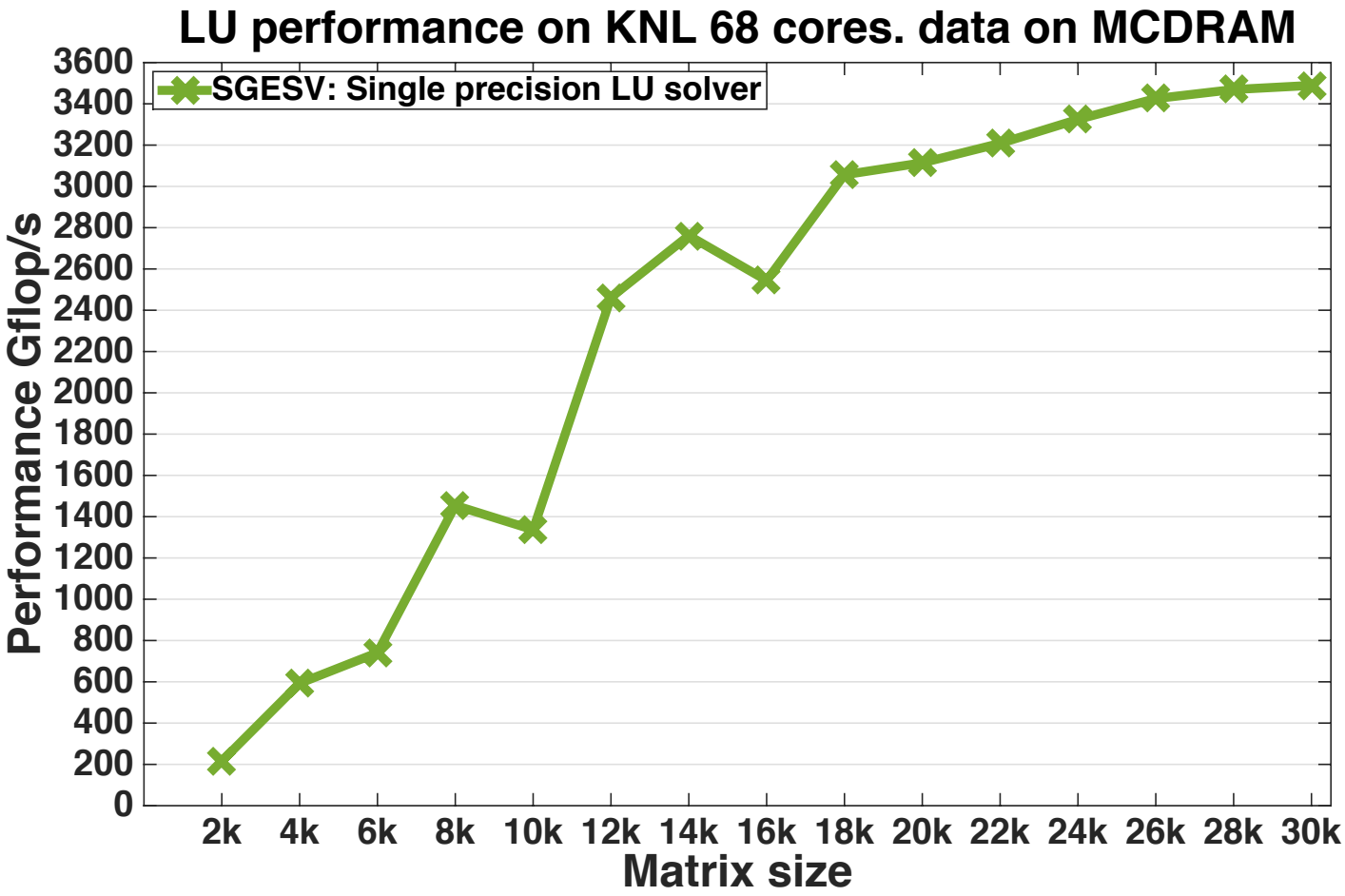
L U = lu(A)	SINGLE	$O(n^3)$
x = L\(U\b)	SINGLE	$O(n^2)$
r = b – Ax	DOUBLE	$O(n^2)$
WHILE || r || not small enough		
z = L\(U\r)	SINGLE	$O(n^2)$
x = x + z	DOUBLE	$O(n^1)$
r = b – Ax	DOUBLE	$O(n^2)$
END		

- Wilkinson, Moler, Stewart, & Higham provide error bound for SP fl pt results when using DP fl pt.
- It can be shown that using this approach we can compute the solution to 64-bit floating point precision.

- Requires extra storage, total is 1.5 times normal;
- $O(n^3)$ work is done in lower precision
- $O(n^2)$ work is done in high precision
- Problems if the matrix is ill-conditioned in sp; $O(10^8)$

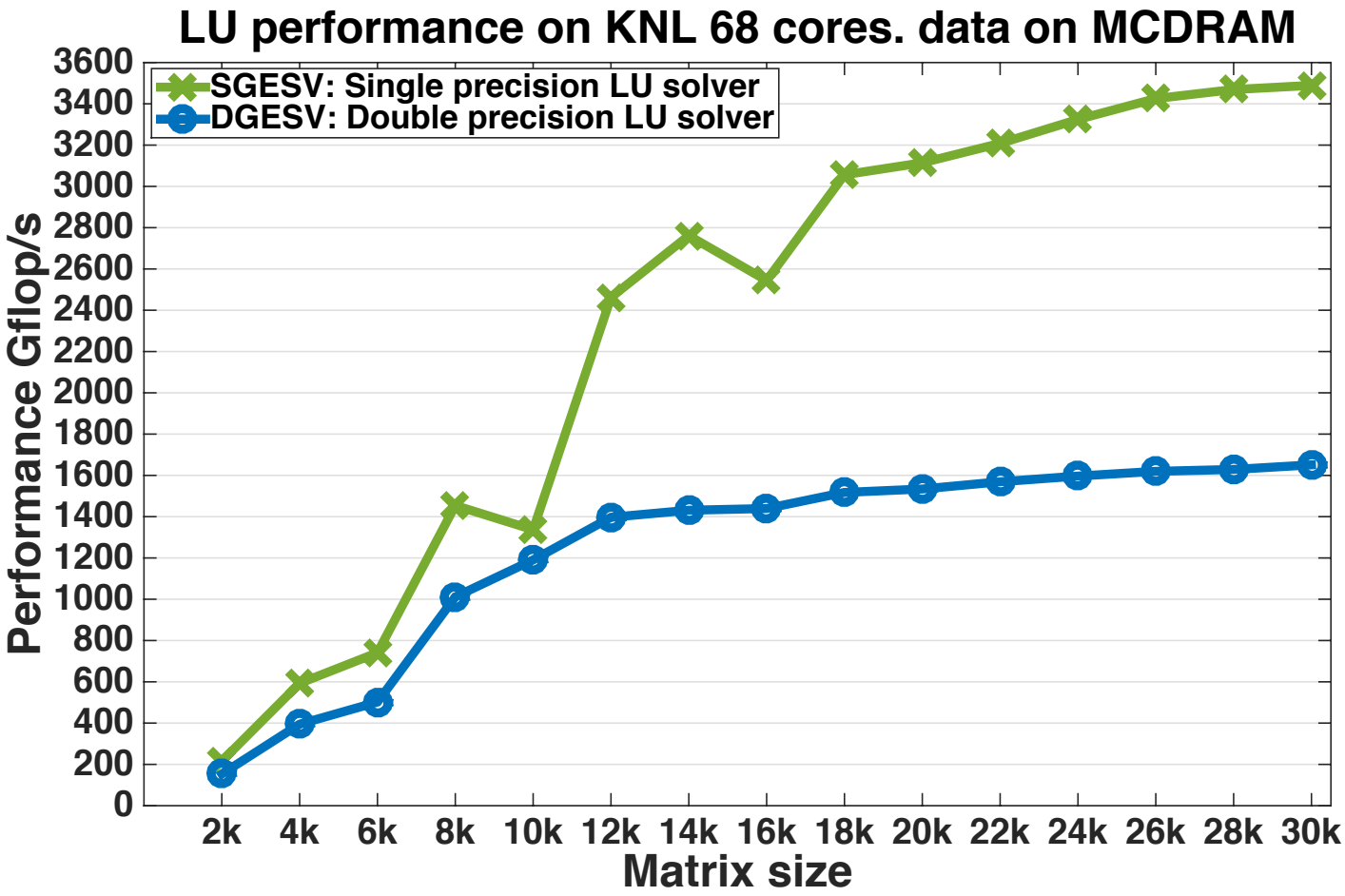
Power-awareness in Algorithms

Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



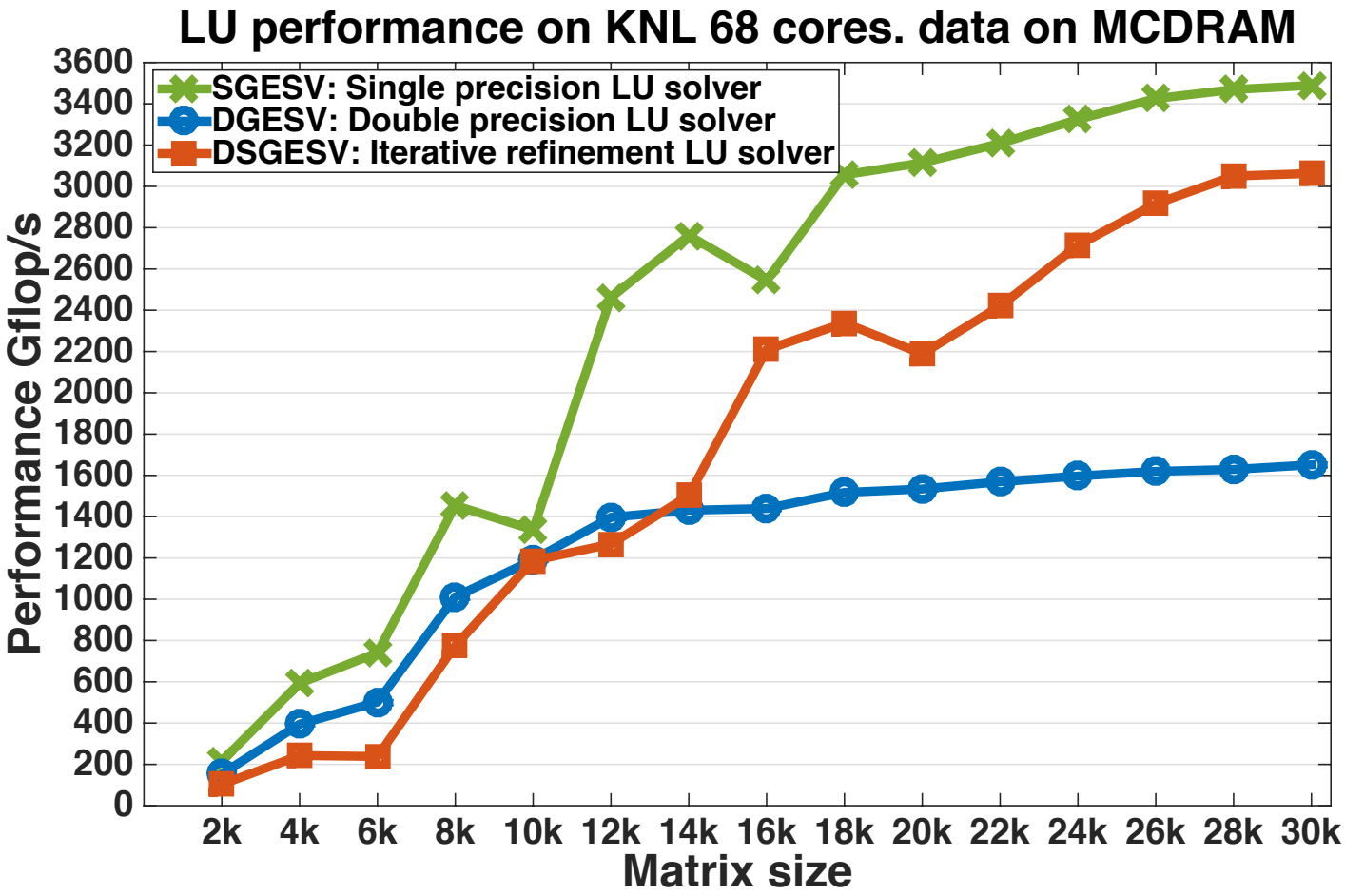
Power-awareness in Algorithms

Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



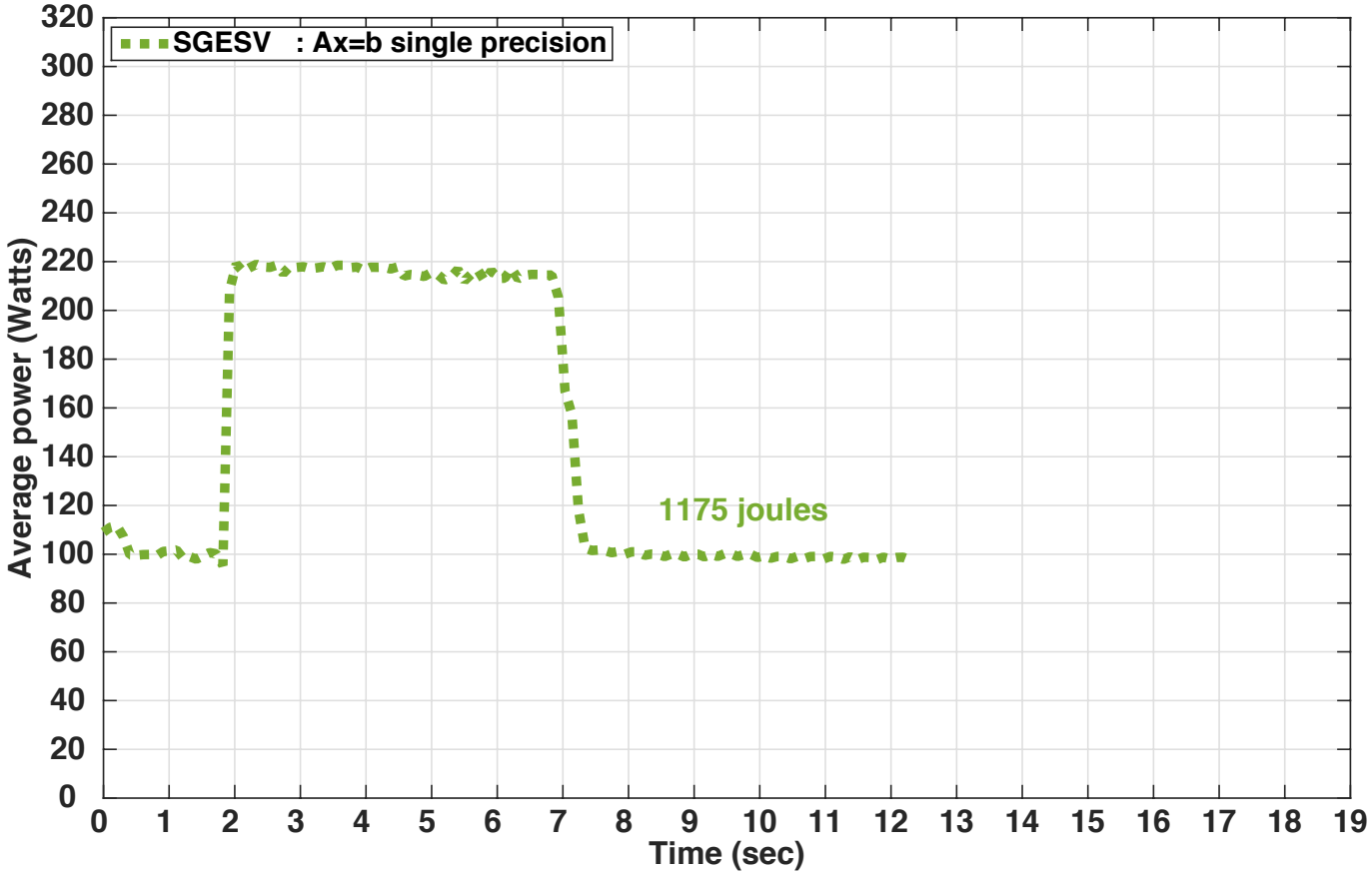
Power-awareness in Algorithms

Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



Power-awareness in Algorithms

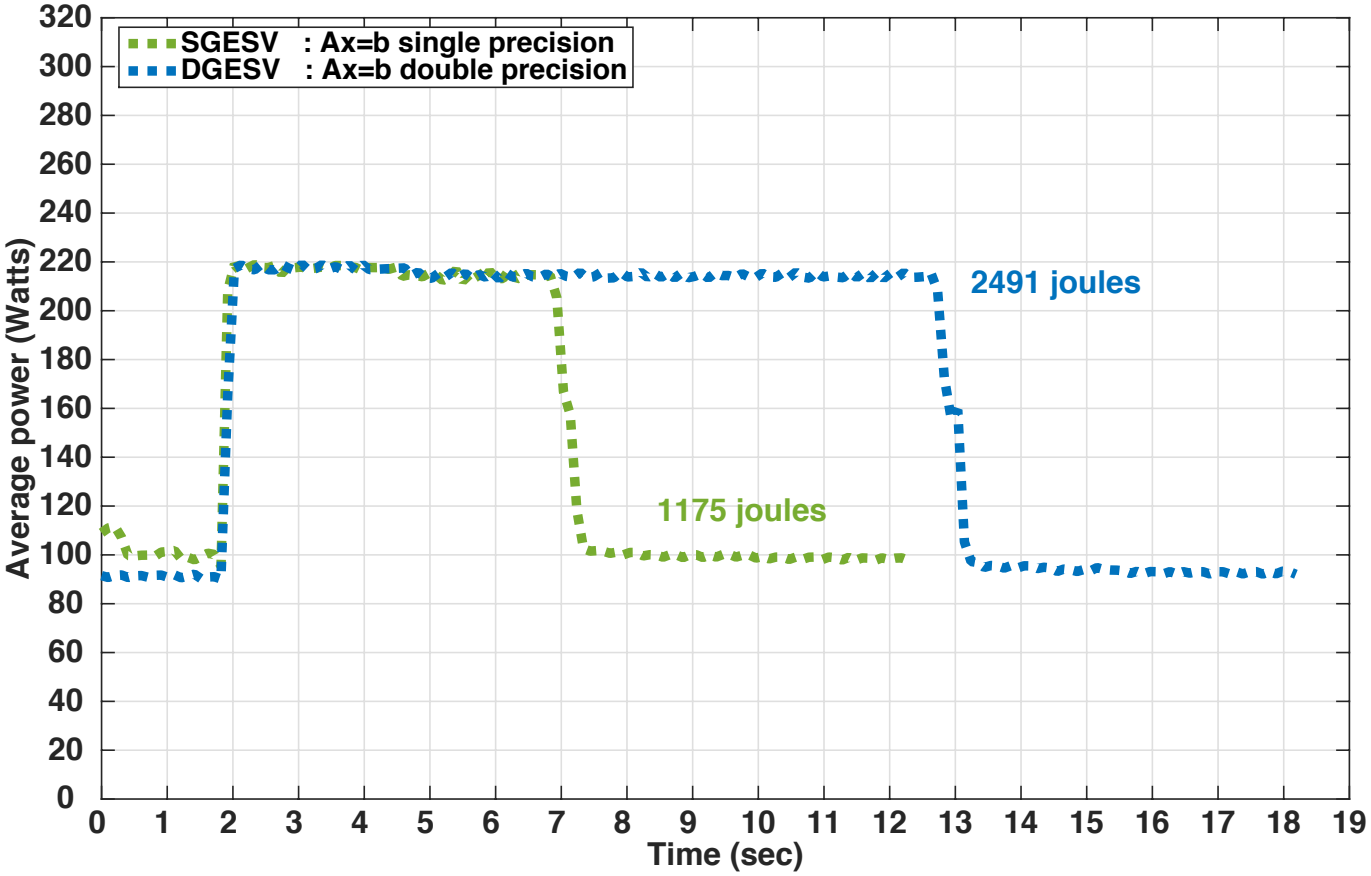
Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



- Power consumption of SP, DP, and mixed precision algorithm to solve Ax=b for a matrix of size 30K on KNL.

Power-awareness in Algorithms

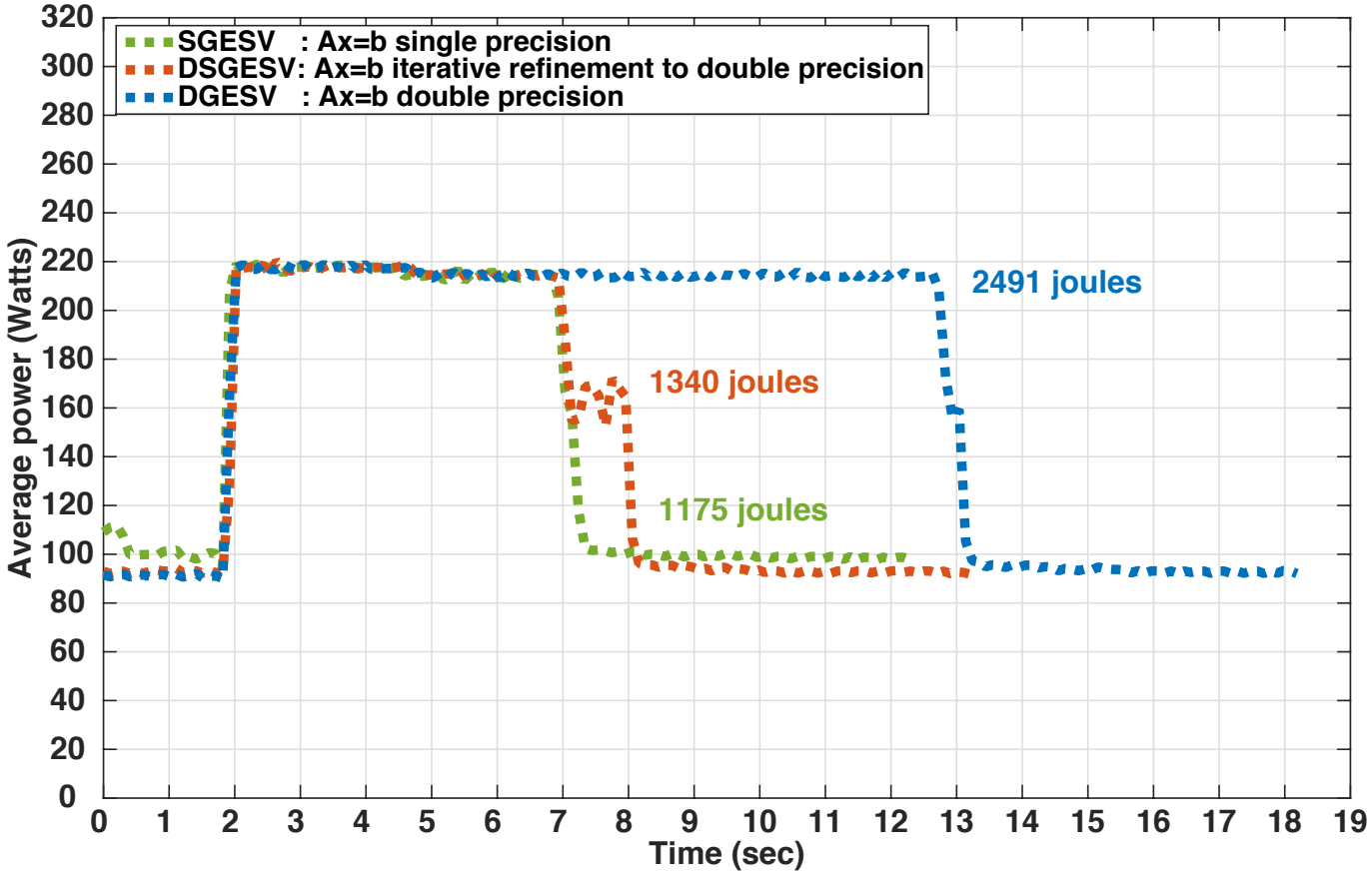
Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



- Power consumption of SP, DP, and mixed precision algorithm to solve Ax=b for a matrix of size 30K on KNL.

Power-awareness in Algorithms

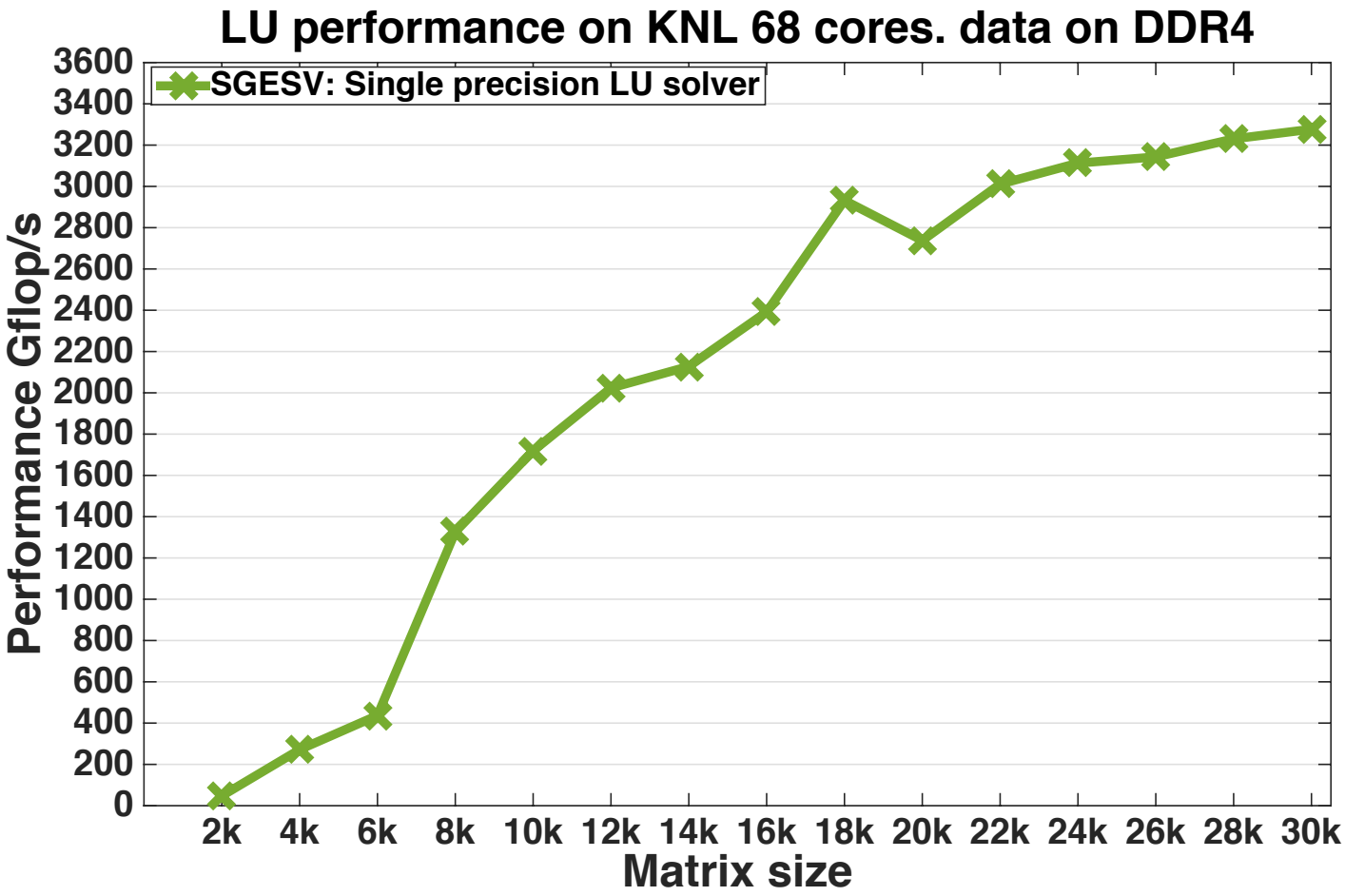
Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



- Power consumption of SP, DP, and mixed precision algorithm to solve Ax=b for a matrix of size 30K on KNL.
- Algorithmic advancements such as mixed precision techniques can also provide a large gain in energy and power consumption. We can reduce the energy by about the half.

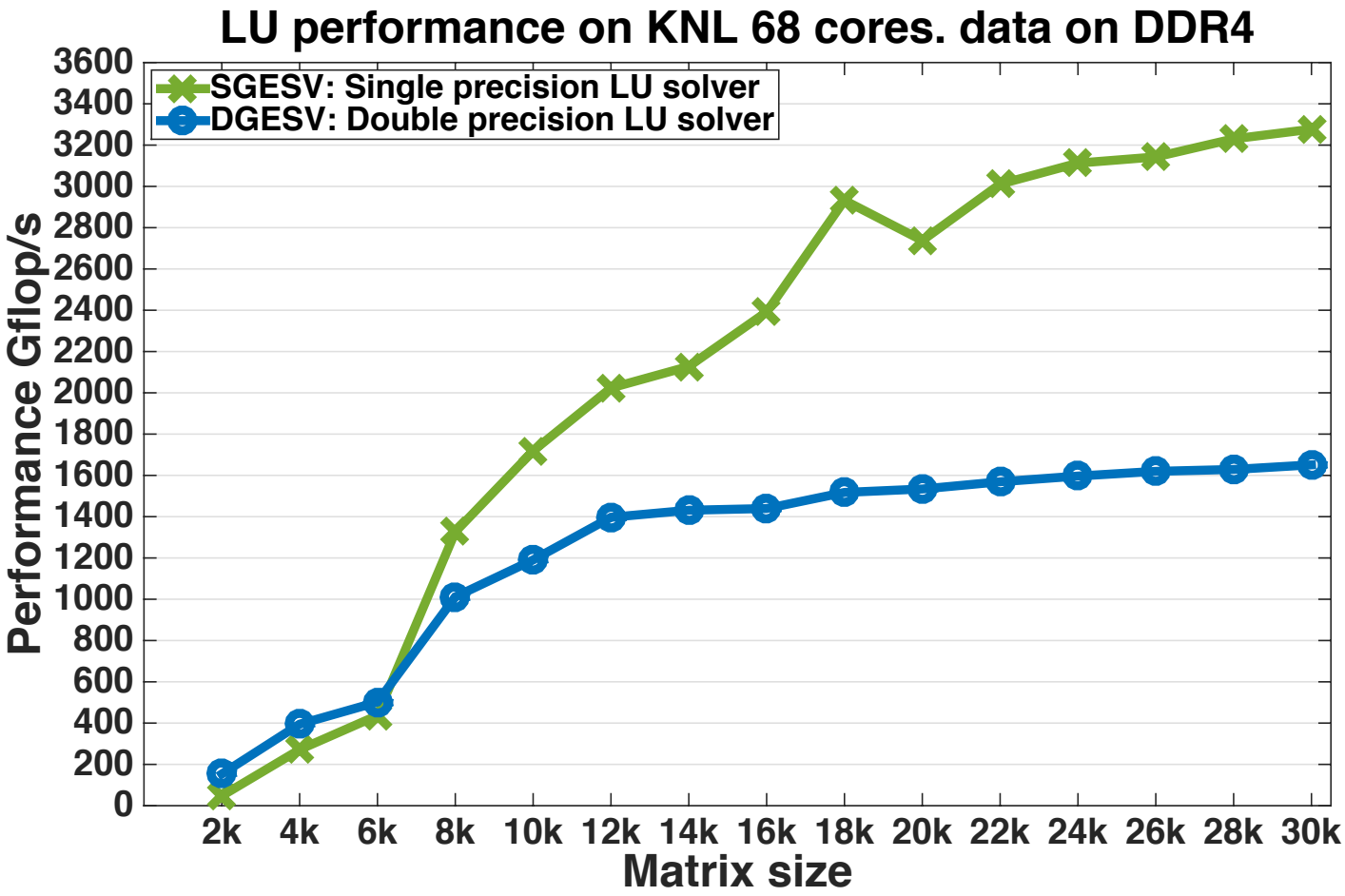
Power-awareness in Algorithms

Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



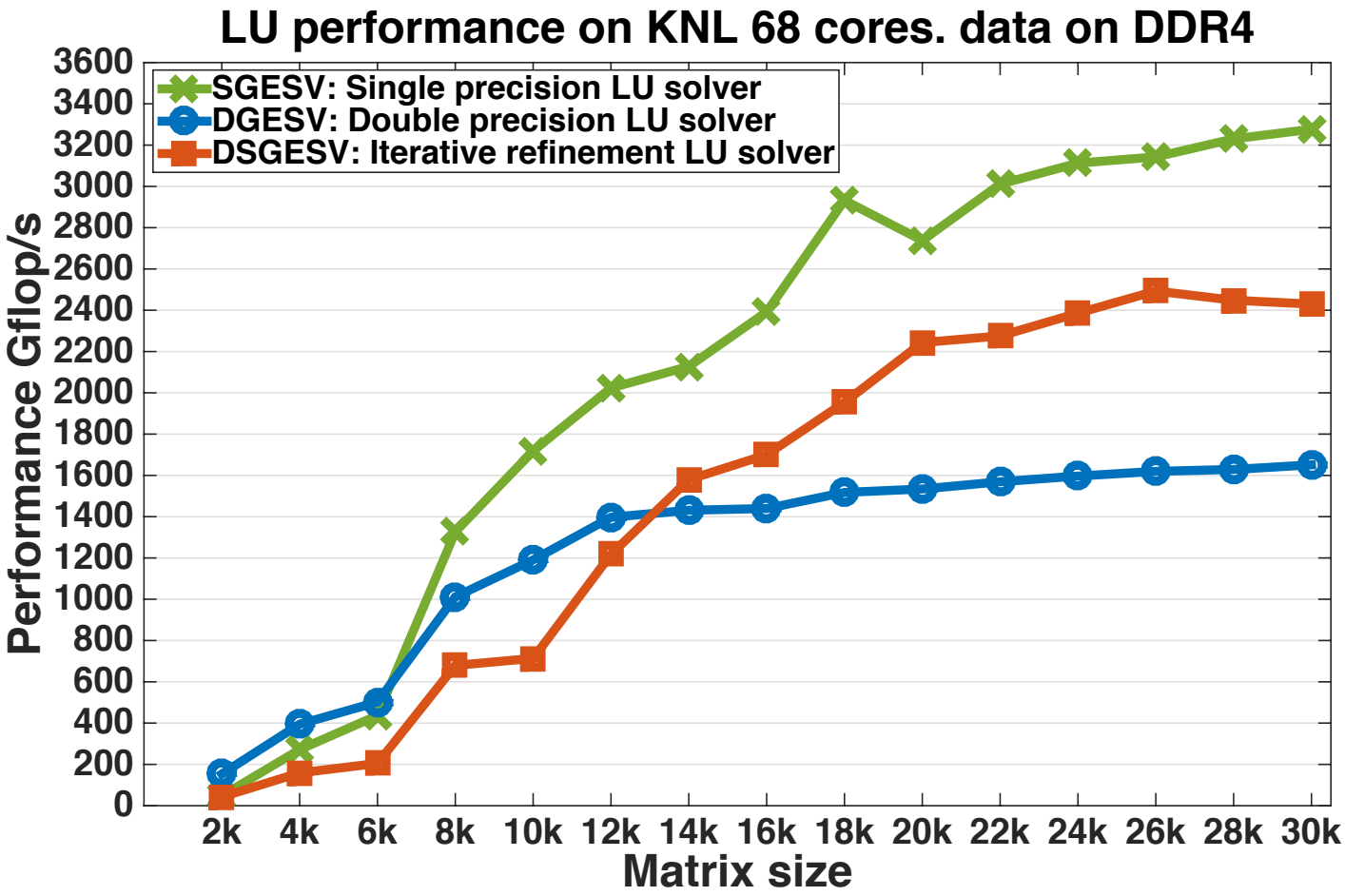
Power-awareness in Algorithms

Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



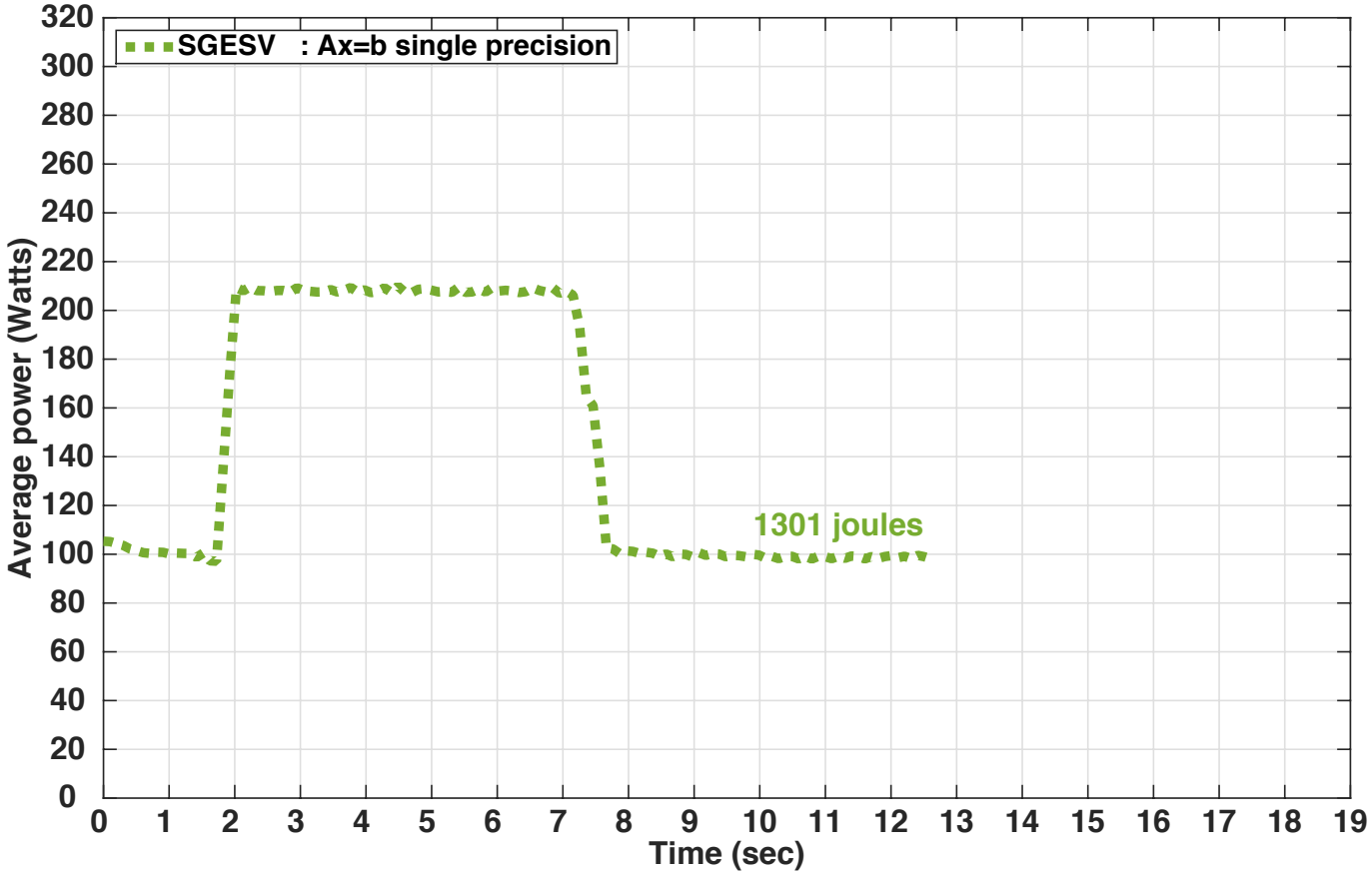
Power-awareness in Algorithms

Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



Power-awareness in Algorithms

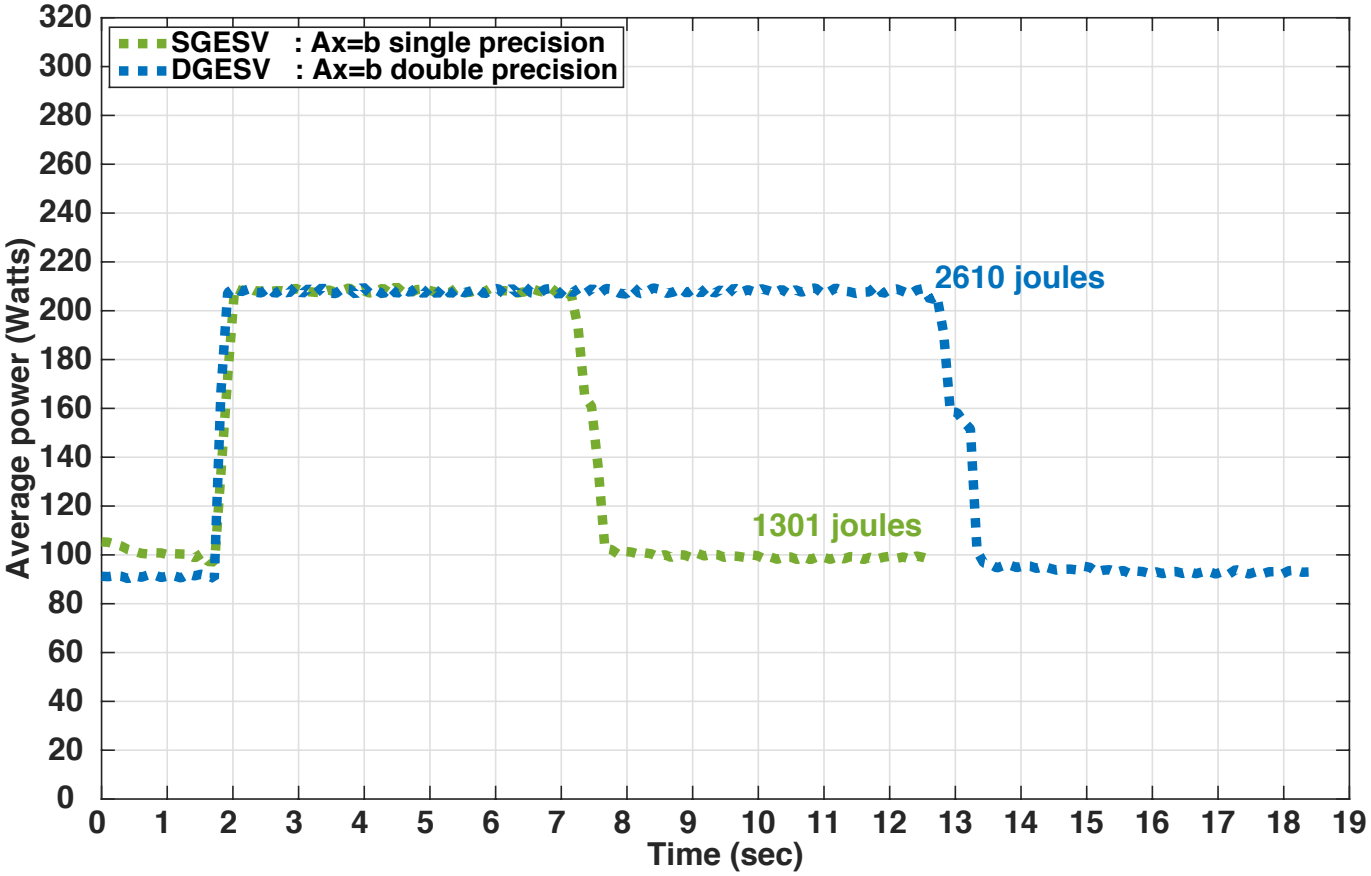
Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



- Power consumption of SP, DP, and mixed precision algorithm to solve Ax=b for a matrix of size 30K on KNL.

Power-awareness in Algorithms

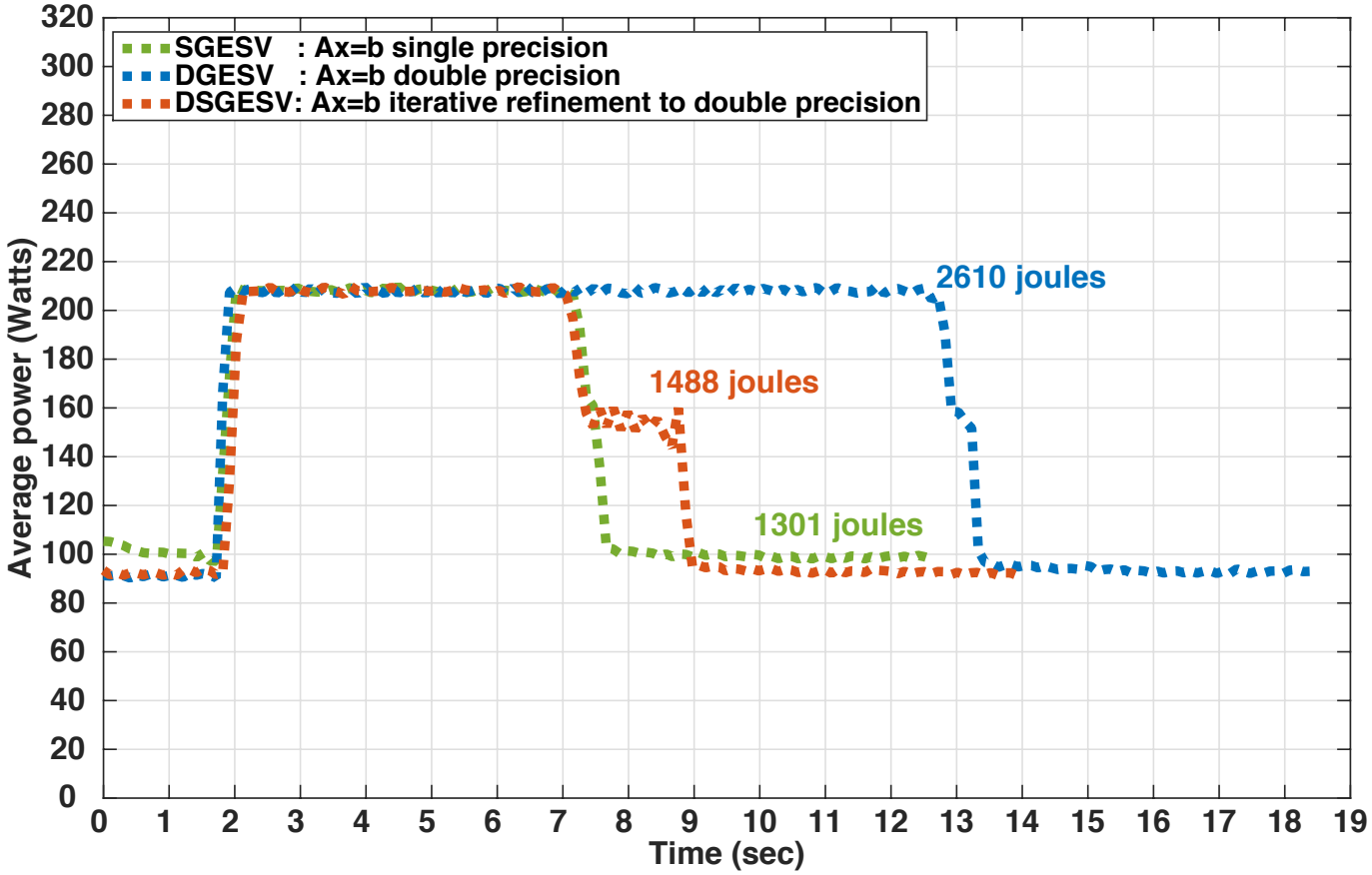
Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



- Power consumption of SP, DP, and mixed precision algorithm to solve Ax=b for a matrix of size 30K on KNL.

Power-awareness in Algorithms

Iterative refinement to solve Ax=b getting a solution in double precision arithmetic



- Power consumption of SP, DP, and mixed precision algorithm to solve Ax=b for a matrix of size 30K on KNL.
- Algorithmic advancements such as mixed precision techniques can also provide a large gain in energy and power consumption. We can reduce the energy by about the half.

3rd Party Tools applying PAPI

- PaRSEC (UTK) http://icl.cs.utk.edu/parsec/
- TAU (U Oregon) http://www.cs.uoregon.edu/research/tau/
- PerfSuite (NCSA) http://perfsuite.ncsa.uiuc.edu/
- HPCToolkit (Rice University) http://hpctoolkit.org/
- Score-P http://score-p.org/
- SCALASCA (FZ Juelich, TU Darmstadt) http://scalasca.org/
- VampirTrace and Vampir (TU Dresden) http://www.vamir.eu
- Open|Speedshop (SGI) http://oss.sgi.com/projects/openspeedshop/
- SvPablo (RENCI at UNC) http://www.renci.org/research/pablo/
- ompP (UTK) http://www.ompp-tool.com

Conclusions and Future Work

- Power efficiency designs – challenges for power and energy-awareness for numerical libraries
- PAPI-based tools to understand and control power usage through power capping
- Characterize algorithms (from BLAS and benchmarks) and quantify possible benefits from power capping on Intel KNL processors

- The tools and analyses presented can be build into libraries to automatically control power & energy consumption based on feedback from computation (at runtime)

Collaborators and Support

MAGMA team

http://icl.cs.utk.edu/magma

Collaborating partners

University of Tennessee, Knoxville
University of Manchester, Manchester, UK
University of Paris-Sud, France
Lawrence Livermore National Laboratory, Livermore, CA
University of California, Berkeley
University of Colorado, Denver
INRIA, France (StarPU team)
KAUST, Saudi Arabia

Intel Parallel Computing Centers

Umeå University
INRIA
Rutherford Appleton Laboratory
University of Manchester