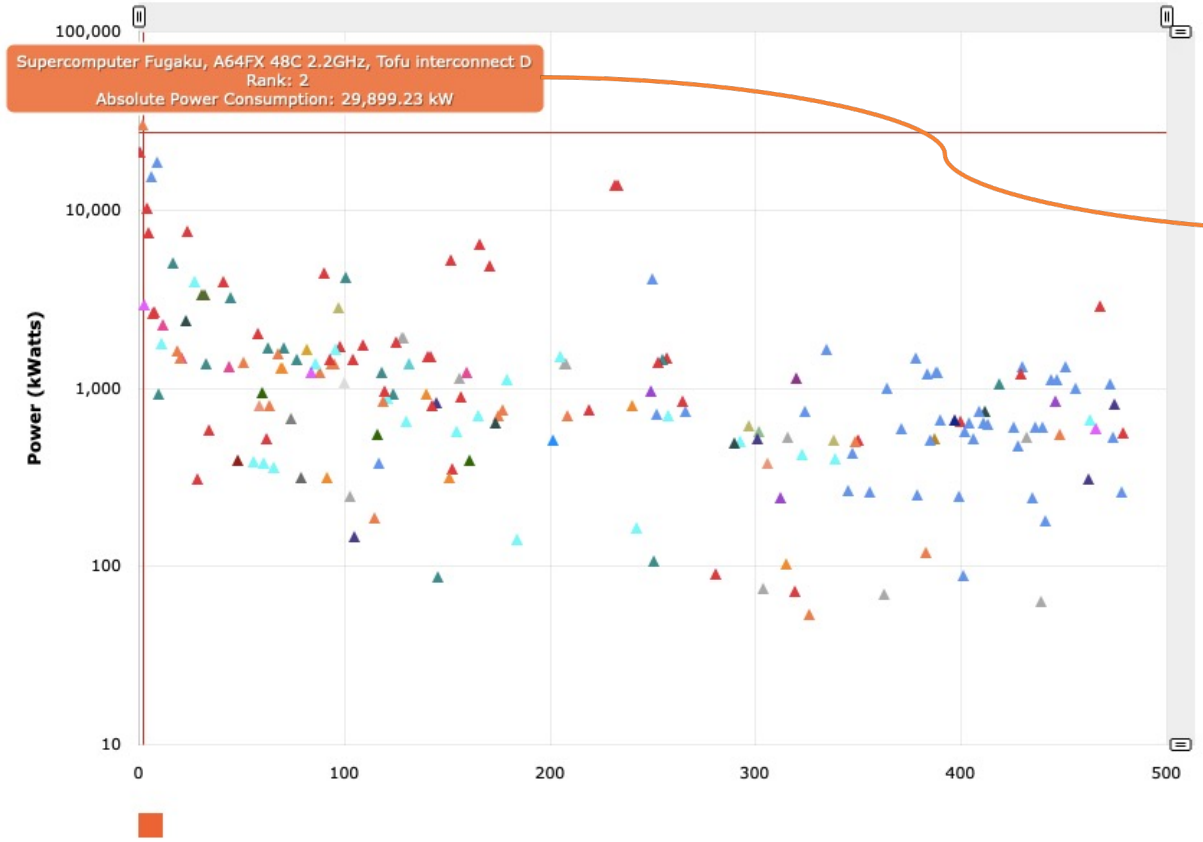
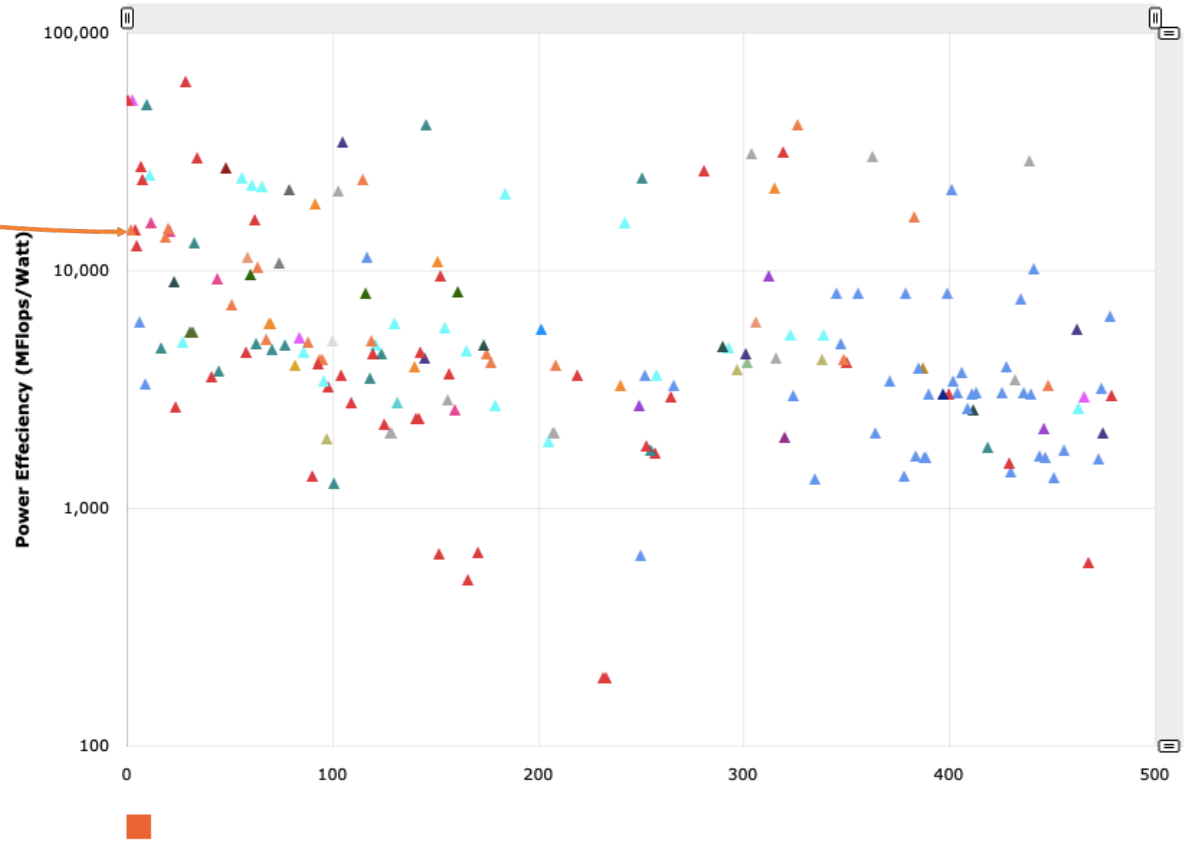
System for collecting statistics on power consumption of supercomputer applications

E.A. Kiselev, A.V. Baranov, P.N. Telegin, E.E. Kuznetsov

Joint Supercomputer Center of the RAS, Russia

The impact of increasing power on the scalability of computing systems





Legend:

Japan, China, United States, Germany, Ireland, France, Australia, United Kingdom, Brazil, Netherlands, Saudi Arabia, Russia, South Korea, Canada, Sweden, Luxembourg, Taiwan, Austria, Poland, Switzerland, Norway, India, Singapore, Italy, Hungary, Czechia, Slovenia, Morocco, Bulgaria, Finland, Spain, United Arab Emirates,

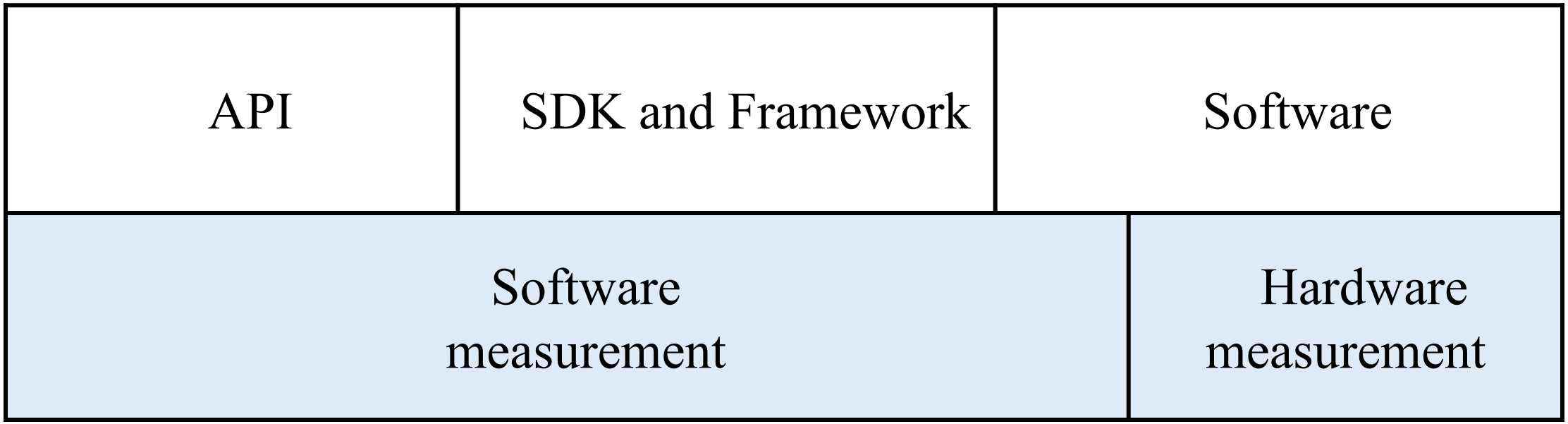
Legend:

Japan, China, United States, Germany, Ireland, France, Australia, United Kingdom, Brazil, Netherlands, Saudi Arabia, Russia, South Korea, Canada, Sweden, Luxembourg, Taiwan, Austria, Poland, Switzerland, Norway, India, Singapore, Italy, Hungary, Czechia, Slovenia, Morocco, Bulgaria, Finland, Spain, United Arab Emirates,

The impact of increasing power on computer system failures

- Calculation results loss during node shutdown or reboot
- Power outage due to overload of the electrical power network
- Increasing job queue time due to computing nodes locking
- Higher frequency of failures requires using more checkpoints, and this results in increased job execution time

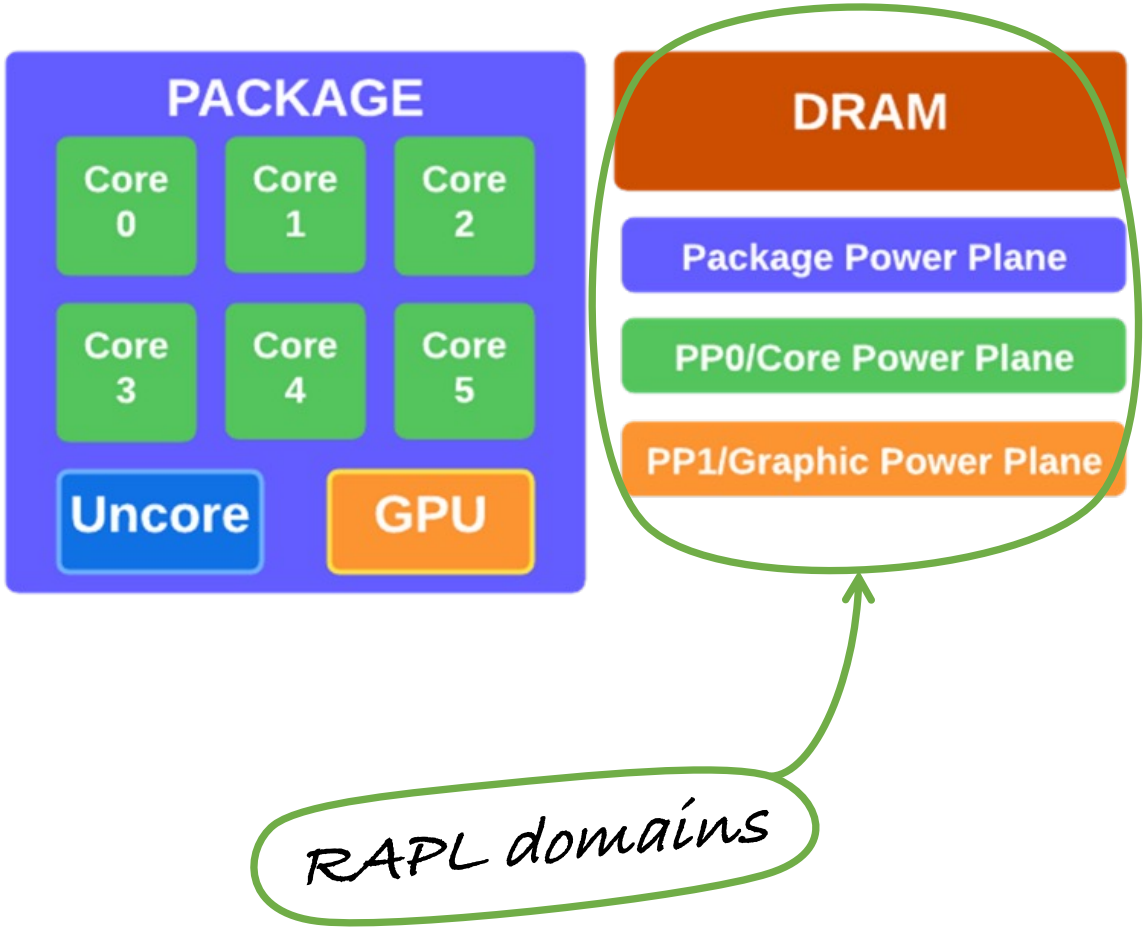
Methods of power consumption measuring in HPC systems



Software and hardware approaches for microprocessors power consumption measurement

API	SDK and Framework	Software	
Software measurement			hardware measurement
PowerAPI, PAPI, Hwmon, perf_events, perfmon, HPM	Intel Energy Checker, Power Capping Framework (powercap)	pTop, PowerTop, Joulemeter	PowerScope, Powermeter, PowerPack, PowerMon2, PowerInsight

RAPL-based approaches for microprocessors power consumption measurement



Running Average Power Limit (RAPL) microcode applies a power consumption prediction model basing on the data acquired from the hardware performance counters

Linux-compatible RAPL-based power consumption measurement:

1. Linux kernel msr-drivers (Intel and AMD comp.)
2. Sysfs virtual file system via powercap-interface (Intel comp.)
3. Software interfaces
    1. perf_event (Intel comp.)
    2. hwmon (Intel and AMD comp.)

RAPL domains supported by different generations of Intel and AMD microprocessors

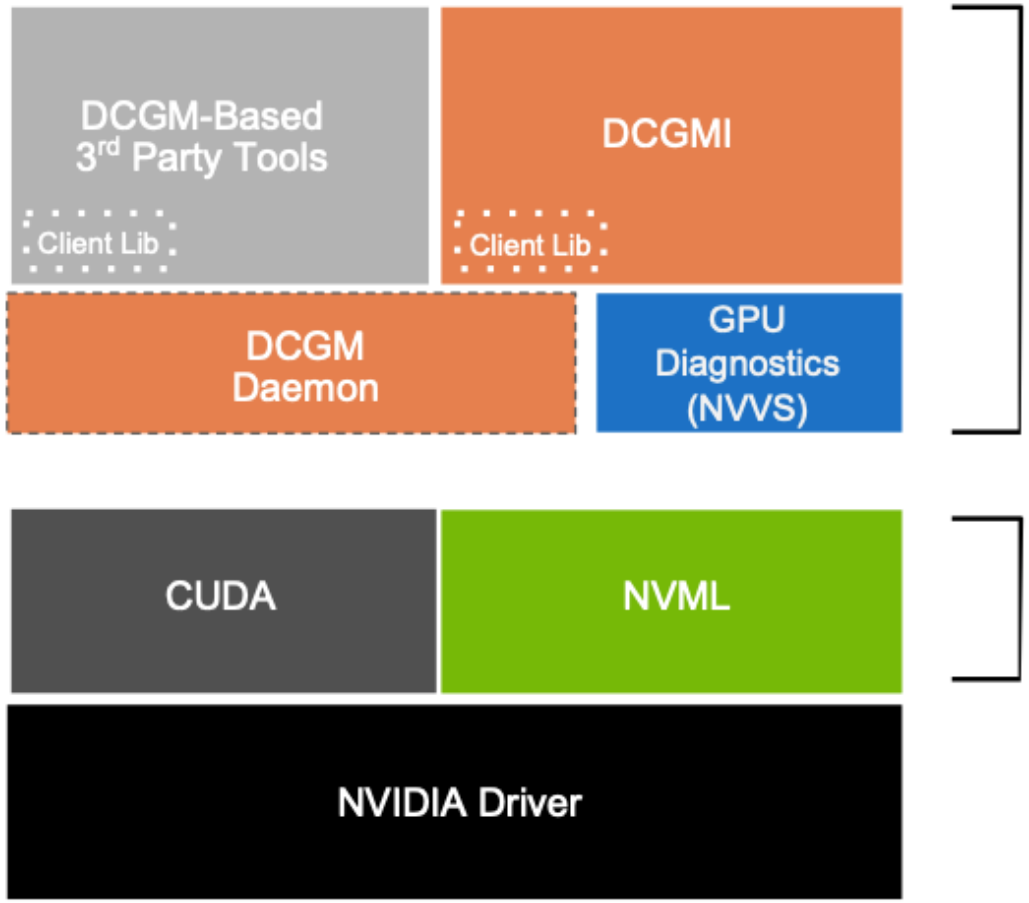
Name	Package domain	PP0	PP1	DRAM
Intel Sandy Bridge	+	+	+	-
Intel Ivy Bridge	+	+	+	-
Intel Haswell	+	+	+	+
Intel Broadwell	+	+	+	+
Intel Skylake	+	+	+	+
Intel Kaby lake	+	+	+	+
Intel Cascade lake	+	+	+	+
Intel Knights Landing	+	-	-	+
Intel Knights Mill	+	-	-	+

Name	Package domain	PP0	PP1	DRAM
AMD Bulldozer	-	-	-	-
AMD Piledriver	-	-	-	-
AMD Piledriver/Trinity	-	-	-	-
AMD Steamroller "Kaveri"	-	-	-	-
AMD Excavator "Carrizo"	-	-	-	-
AMD Jaguar "Mullins"	-	-	-	-
AMD Zen	+	+	-	-
AMD Zen+	+	+	-	-
AMD Zen2	+	+	-	-

Methods of power consumption measuring for graphics microprocessors

Nvidia	AMD
1. NVML API (nvidia-smi) 2. DCGM API	• lm-sensor

Available NVIDIA management tools



Data Center GPU Manager (DCGM)

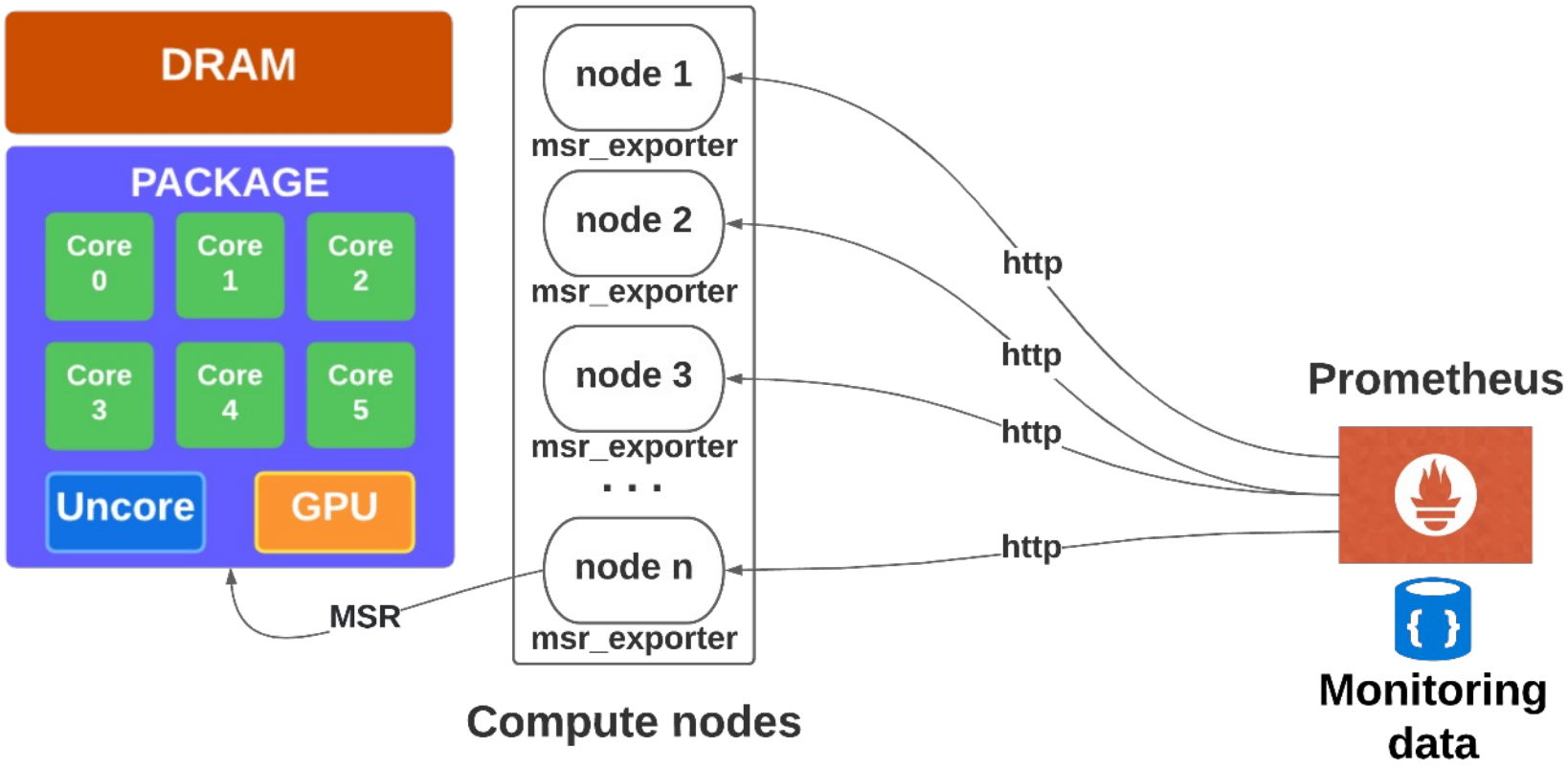
- Additional diagnostics (aka NVVS) and active health monitoring
- Policy management and more

NVIDIA Management Library (NVML)

- Low level control of GPUs
- Included as part of driver
- Header is part of CUDA Toolkit / DCGM

NVIDIA

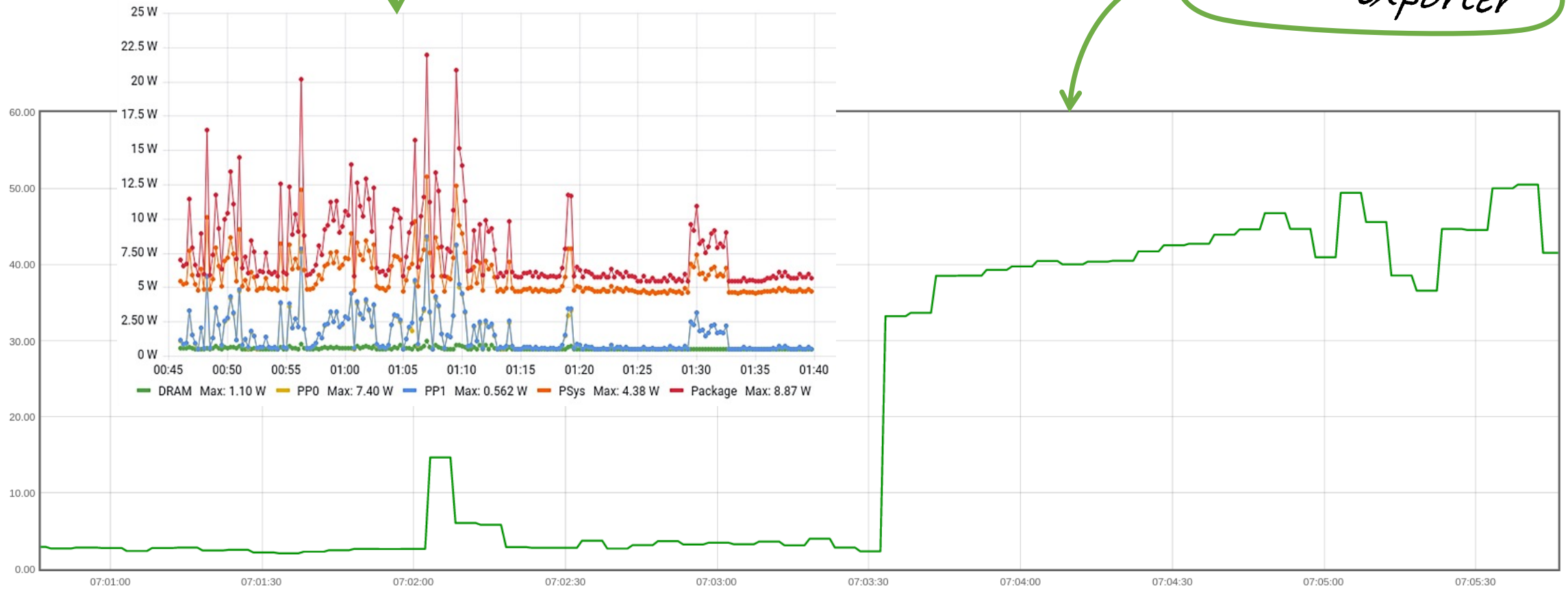
Software stack for collecting data on computers power consumption



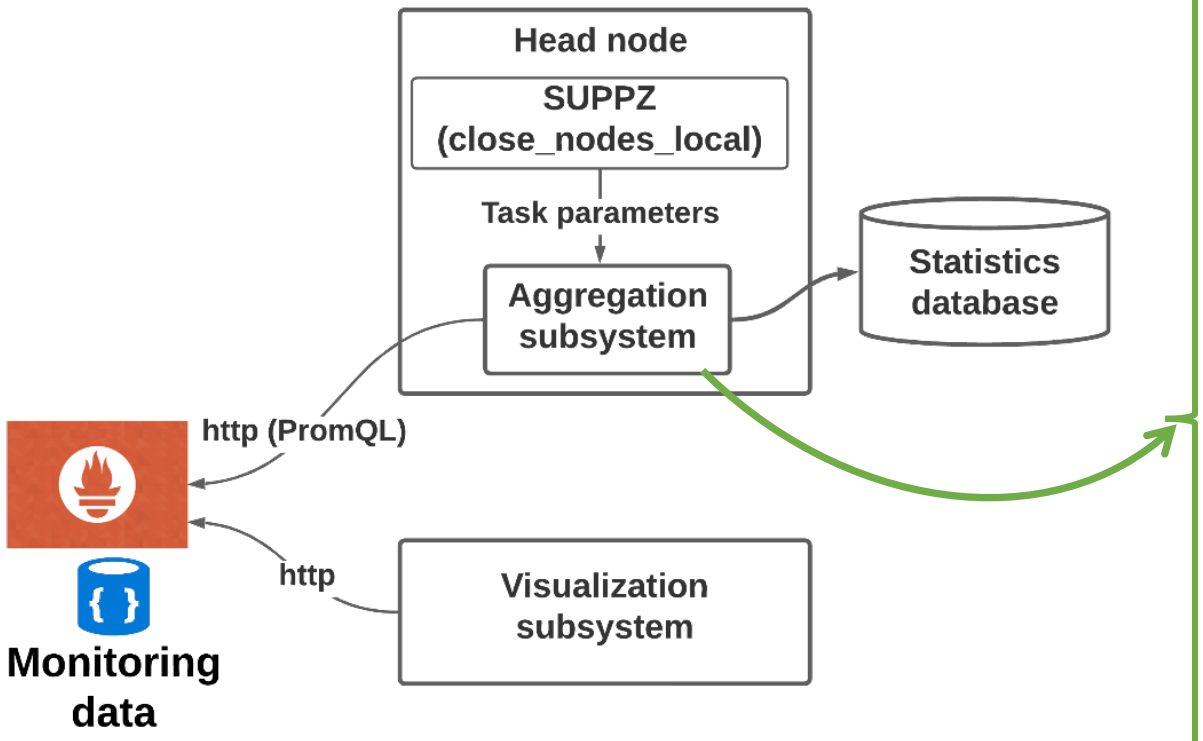
Prometheus exporters for Intel and Nvidia GPU power consumption data collection

msr-exporter

nvml-exporter



Software stack for collecting data on computers power consumption



- E_{max} (J/s) – maximum total power consumption of all allocated computing nodes during job runtime
- E_{med} (J/s) – median value of allocated nodes power consumption:

$$E_{med} = x'_{(n+1)/2} \text{ for odd n,}$$

$$E_{med} = \frac{x'_{n/2} + x'_{(n+1)/2}}{2} \text{ for even n.}$$

- E_{sum} (J) – the total value of power consumption of all allocated nodes during a job runtime
- $C_{med}(U)$ (cores number) – median value of the number of cores with load per cent exceeding the set limit value U
- U_{med} (%) – median value of the computing node cores load percent

NPB test jobs run parameters

Test name	Required number of cores	Number of allocated computing nodes			
		Broadwell	Cascade lake	KNL	Skylake
BT	144	5	3	2	4
EP	144	5	3	2	4
IS	256	8	6	4	8
LU	256	8	6	4	8
SP	256	8	6	4	8

Experimental results

MVS10P OP partition	BT	EP	IS	LU	SP
E_{max} (J/s)					
Broadwell	1257.3	1242.1	1665.8	2259.6	2101.6
Cascade lake	1407.7	1327.7	2259.1	2437,1	2582.9
Skylake	1698.9	1668.5	2447.1	2949.5	3233.3
KNL	417.5	308.1	622.4	727.3	792.1
E_{med} (J/s)					
Broadwell	1234.4	1237.7	1645.6	2243.8	2080.7
Cascade lake	1336.8	1274.3	2008.7	2416.2	2555.6
Skylake	1683.2	1656.1	2398.5	2937.6	3206.1
KNL	397.8	306.8	599.9	709.6	775.2

MVS10P OP partition	BT	EP	IS	LU	SP
E_{sum} (J)					
Broadwell	265769	30765.9	15783.5	172417.2	335122.6
Cascade lake	227116.2	31585.9	17115.3	172090.7	294234.6
Skylake	282044.1	37303.3	13217.9	188012.6	364193.2
KNL	85568.4	39289.4	7765.5	222308.9	235798.8
C_{med} (number of cores)					
Broadwell	144	144	236	256	256
Cascade lake	144	144	179	256	256
Skylake	144	144	163	256	256
KNL	144	144	192	254	255
U_{med} (%)					
Broadwell	100	100	99.8	100	100
Cascade lake	100	100	78.3	100	100
Skylake	100	100	96,3	100	100
KNL	100	100	93.7	99.6	99.8

Experimental results

- The obtained values show that one and the same algorithm can have different impact on different computers power consumption
- The difference between the maximum E_{max} and the median E_{med} values of power consumption is not always related to the higher load of the processor cores
- The results of comparing E_{max}, E_{med}, U_{med} and C_{med} values demonstrate the need to examine them in a bundle to determine the impact of a user application on energy consumption
- It can be assumed that there is a computation time limit at which less productive, but energy-efficient systems can reduce total energy consumption
- It is notable that the values of E_{max}, E_{med}, U_{med} and C_{med} allow for the conclusion of the workload character during a parallel algorithm execution

Conclusion

- To solve the problem of power consumption statistics collecting, processing, and accounting at supercomputer jobs runtime, the needed software stack, as well as the list of statistical indicators necessary for parallel applications power consumption description were determined
- The results show the practicability of accounting and control of the impact that parallel applications execution has on computers power consumption
- Controlling the impact of parallel programs execution on the energy consumption enables both tracking instantaneous and peak power consumption loads in computers and analyzing the statistics of using the computing resources by the users to identify user jobs energy profiles
- The information on the energy profiles allows implementing energy efficient job scheduling at a supercomputer center