Green Computing

v praxi

Lukáš Hejtmánek, CERIT-SC, e-INFRA CZ

Linux Days 2024

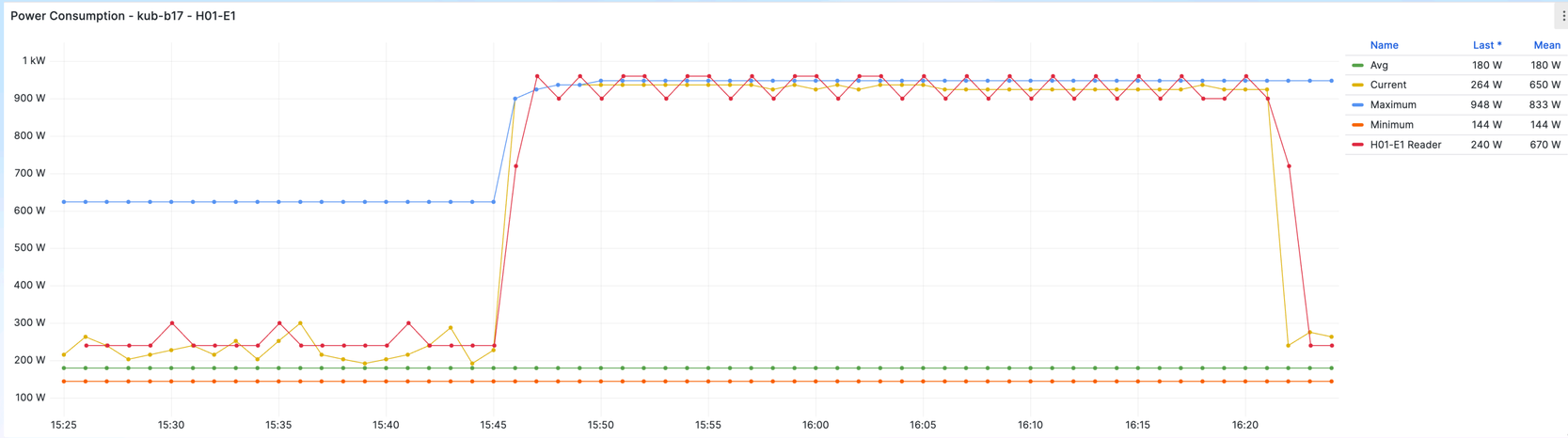
Projekty Fondu Rozvoje — CESNET

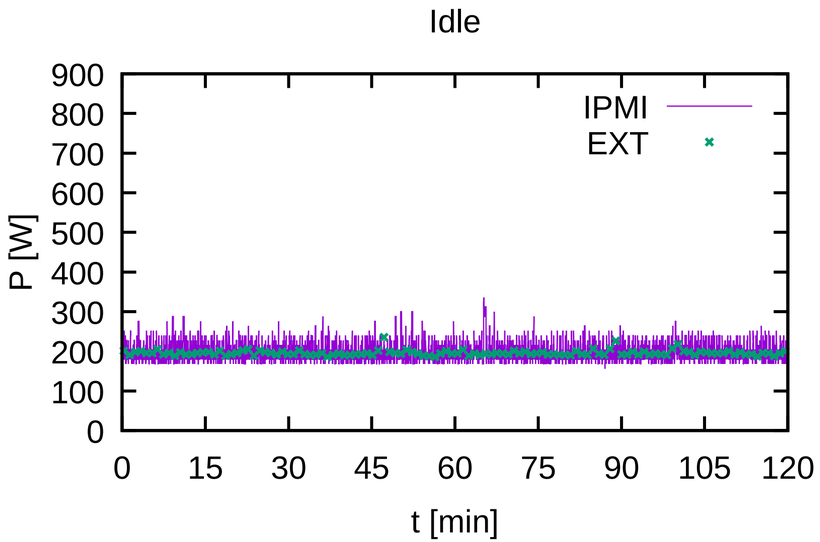
Green computing in Academic Datacenter

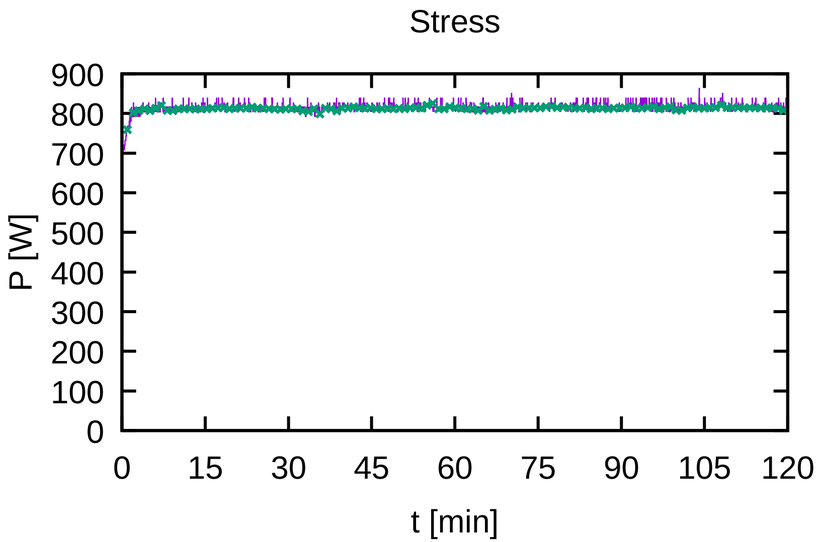
- Cíl projektu: Jaké jsou opravdové možnosti úspory energie
- Reálná měření v datacentru CERIT-SC Masarykovy Univerzity v Brně
- Úspora energie pro
  - CPU
  - GPU
  - úložiště
  - chlazení serverů

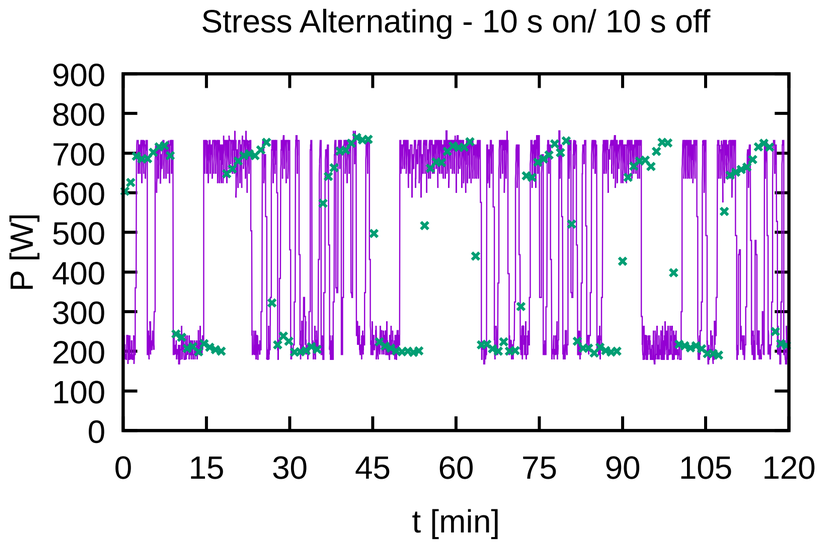
Kalibrace měření spotřeby

- Jak měřit spotřebu serverů?
  - Interní sensory — IPMI
    - Jak jsou přesné?
    - Bez integrace v čase
  - Externí měřiče spotřeby
    - Přesné měření
    - Malé rozlišení (problém pro malé odběry)
    - S integrací v čase




Idle
IPMI
EXT
P [W]
t [min]
0
100
200
300
400
500
600
700
800
900
0
15
30
45
60
75
90
105
120


Stress
P [W]
t [min]
0
100
200
300
400
500
600
700
800
900
0
15
30
45
60
75
90
105
120


Stress Alternating - 10 s on/ 10 s off
P [W]
t [min]
0
100
200
300
400
500
600
700
800
900
0
15
30
45
60
75
90
105
120

Regulace CPU výkonu

- Většina CPU dovoluje měnit frekvence
  - Frekvence  spotřeba
  - Regulace frekvence v Linuxu
    - scaling governor
      - Powersave
      - Performance
      - Schedutil
      - Schedutil 2GHz

Regulace CPU výkonu

- IDLE

Governor	IPMI	IPMI Relative	External Source	External Relative
Schedutil	240 Watts	100.0%	254 Watts	100.0%
Powersave	237 Watts	98.6%	253 Watts	99.6%
Performance	246 Watts	102.5%	259 Watts	102.0%

- Busy

Governor	IPMI	IPMI Relative	External Source	External Relative
Schedutil	930 Watts	100.0%	931 Watts	100.0%
Powersave	349 Watts	37.5%	354 Watts	38.0%
Performance	930 Watts	100.0%	931 Watts	100.0%

- Alternating

Governor	IPMI	IPMI Relative	External Source	External Relative
Schedutil	569 Watts	100.0%	586 Watts	100.0%
Powersave	292 Watts	51.3%	304 Watts	51.9%
Performance	579 Watts	101.8%	593 Watts	101.2%

Regulace CPU výkonu

Reálné aplikace

- Nextflow pipeline

Governor	Run Time	Runtime Relative	External Source	External Relative
Schedutil	9.51 Hours	100.0%	478 Watts	100.0%
Schedutil 2G	16.42 Hours	172.6%	336 Watts	70.3%
Powersave	20.78 Hours	218.5%	318 Watts	66.5%
Performance	9.64 Hours	101.4%	485 Watts	101.5%

- SPEC CPU2017

Governor	Score	Relative Score	External Source	External Relative
Schedutil	323	100.0%	829 Watts	100.0%
Schedutil 2G	205	63.5%	449 Watts	54.2%
Powersave	155	48.0%	397 Watts	48.0%
Performance	321	99.4%	832 Watts	100.4%

Regulace GPU výkonu

- Pouze NVIDIA GPU (jiné nemáme k dispozici)
- Dovolují měnit frekvenci jader a pamětí
  - nvidia-smi
- IDLE spotřeba poměrně vysoká
  - 70W na H100 NVL kartě
  - Pozor na běžící monitoring karty, zvyšuje IDLE spotřebu (dcgm)

Regulace GPU výkonu

- Compute performance

Test type	<6251, 1695>	<6251, 210>	<405, 420>	<405, 210>
sha512crypt 6, SHA512 (Unix)	98555 H/s	27201 H/s	57092 H/s	26689 H/s
KeePass 1 (AES/Twofish) and KeePass 2 (AES)	97679 H/s	13898 H/s	27810 H/s	13897 H/s
Kerberos 5, etype 23, TGS-REP	935 MH/s	135 MH/s	271 MH/s	135 MH/s
7-Zip	231 kH/s	34 kH/s	72 kH/s	34 kH/s

- Power usage

Clocks combination	Mean W from External Sensor	Mean W from Internal Sensor
<6251, 1695>	527W	490W
<6251, 210>	394W	383W
<405, 420>	404W	390W
<405, 210>	389W	386W

Regulace chlazení serveru

- Některé servery dovolují nastavovat profily chlazení
  - Regulace otáček ventilátorů
  - Různé profily
    - Maximální chlazení
    - Vyvážené chlazení
    - Atd. …
    - Nastavení křivky otáček ventilátorů

Regulace chlazení serveru

- Křivky ventilátorů

Fan Speed	External Source	External Relative	Avg Fan Speed	Avg CPU Temp
Standard	253 Watts	100.0%	5346 RPM	44.1°C
Delayed Start	251 Watts	99.2%	3780 RPM	54.3°C
Full Speed	470 Watts	185.8%	16423 RPM	38.3°C

- BIOS profily

BIOS	Score	Relative Score	External Source	External Relative
Standard	323	100.0%	829 Watts	100.0%
Energy Efficient	276	85.4%	696 Watts	84.0%

Regulace výkonu disků

- Disky vesměs regulaci nemají
  - Automatické uspání
    - V datových úložištích — RAID skupinách — není moc užitečné
- Zaměření se na typy RAID konfigurací
  - Mirror — není náročné na CPU
  - RAID 6 ekvivalent — náročnější na CPU
- Zaměření se na rychlost CPU a dopad na výkon disků

Regulace výkonu disků

- Testovány distribuované souborové systémy
  - GPFS, BeeGFS
  - NVME a rotační disky

Regulace výkonu disků

- GPFS RAID6

Pool setup	Read linear [GB/s]	Write linear [GB/s]	Active power [Watt]
32 nodes HDD	32.4	34.7	301
16 nodes HDD	17.8	19.1	322
32 nodes SSD	46.9	34.8	302
16 nodes SSD	18.0	12.1	322

- GPFS Mirror

Pool setup	Read linear [GB/s]	Write linear [GB/s]	Active power [Watt]
32 nodes HDD	46.8	15.9	272
16 nodes HDD	24.1	9.6	312
32 nodes SSD	48.9	16.4	272
16 nodes SSD	19.4	4.0	319

- BeeGFS Mirror

Pool setup	Read linear [GB/s]	Write linear [GB/s]	Active power [Watt]
32 nodes HDD	16.9	21.9	282
16 nodes HDD	9.1	11.5	285
32 nodes SSD	18.4	20.2	261
16 nodes SSD	10.6	12.2	285

Regulace výkonu disků

Změny CPU frekvence

- GPFS RAID6

Pool and Frequency setup	Read linear [GB/s]	Write linear [GB/s]	Active power [Watt]
HDD 1.5 GHz	35.4	33.6	303
SSD 1.5 GHz	45.8	33.3	274
HDD 2.0 GHz	35.6	34.1	306
SSD 2.0 GHz	46.7	34.0	287
HDD 2.8 GHz	35.1	34.8	310
SSD 2.8 GHz	35.6	34.8	291

- GPFS Mirror

Pool and Frequency setup	Read linear [GB/s]	Write linear [GB/s]	Active power [Watt]
HDD 1.5 GHz	44.5	15.2	308
SSD 1.5 GHz	47.6	16.2	303
HDD 2.0 GHz	45.7	15.9	312
SSD 2.0 GHz	48.7	16.2	290
HDD 2.8 GHz	46.7	15.9	313
SSD 2.8 GHz	48.9	16.4	310

Simulátor cen energie

- Vyvinut jednoduchý simulátor cen energie — zdroj dat OTE
- Myšlenka
  - Drahá energie — snížíme frekvenci CPU ➡️ snížíme spotřebu
  - Levnější energie — zvýšíme frekvenci CPU ➡️ zvýšíme spotřebu
- Simulátor je funkční, frekvence CPU serverů se mění dle aktuální ceny
  - Jen pro demonstrační účely

Výsledky

- Regulace výkonu (spotřeby) funguje
  - Je možné do značné míry regulovat odběr serveru
    - Nejúčinnější je regulace frekvence CPU
    - GPU je komponenta s nejvyšší spotřebou
      - Možnosti regulace spotřeby omezené

Výsledky
Ekonomické dopady

- Snížení spotřeby může ušetřit nemálo peněz za energie
  - Důsledek
    - Snížení využitelnosti HW
    - Cena HW je výrazně vyšší než cena energie
    - Z ekonomického hlediska — regulace spotřeby — pouze pokud musíme
      - Nemáme peníze na další energii
      - Nemáme dostatečné chlazení

Děkuji za pozornost