"Domowy superkomputer"
czyli jak uruchomić Intel® XeonPhi™ pod Linuxem

jesień linuxowa 2015
Hucisko

Marek Sroczyński
mareksr@hm.pl

Cel: Zbudować działający system gotowy do obliczeń.

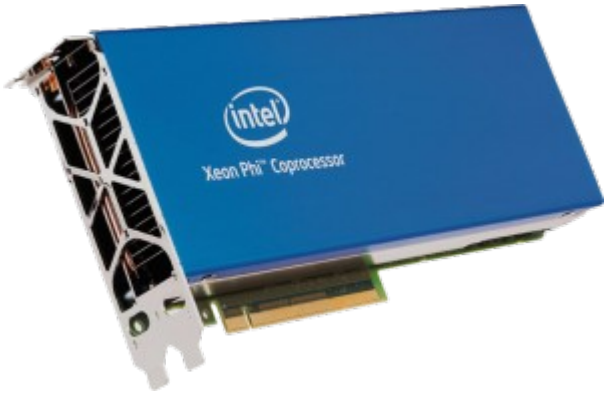
- Architektura Knights Corner
- Hardware dla Host'a
- Instalacja oraz konfiguracja Linux jako Host'a
- Podstawowe narzędzia zarządzania koprocesorem
- Kompilacja offload/natywna
- uOS

ASCI Red:
Sandia National Laboratories

Number 1 system from June 1997 to June 2000

Intel ASCI Red Sandia National Laboratories, USA			
Date	Cores	Linpack Peak	Theoretical Peak
6/97	7,264	1.068 Tflop/s	1.453 Tflop/s
11/97	9,152	1.34 Tflop/s	1.83 Tflop/s
6/98	9,152	1.34 Tflop/s	1.83 Tflop/s
11/98	9,152	1.34 Tflop/s	1.83 Tflop/s
6/99	9,472	2.1 Tflop/s	3.1 Tflop/s
11/99	9,632	2.4 Tflop/s	3.2 Tflop/s
06/00	9,632	2.4 Tflop/s	3.2 Tflop/s
Interconnect: Proprietary Operating System: Paragon OS			

Last appearance on list: No. 276 in November 2005



Product Name	Status	Launch Date	# of Cores	TDP	Recommended Customer Price
Intel® Xeon Phi™ Coprocessor 3120A (6GB, 1.100 GHz, 57 core)	Launched	Q2'13	57	300 W	$1695.00 - $1960.00
Intel® Xeon Phi™ Coprocessor 3120P (6GB, 1.100 GHz, 57 core)	Launched	Q2'13	57	300 W	$1695.00
Intel® Xeon Phi™ Coprocessor 5120D (8GB, 1.053 GHz, 60 core)	Launched	Q2'13	60	245 W	$2759.00
Intel® Xeon Phi™ Coprocessor 5110P (8GB, 1.053 GHz, 60 core)	Launched	Q4'12	60	225 W	$2437.00 - $2649.00
Intel® Xeon Phi™ Coprocessor 7120A (16GB, 1.238 GHz, 61 core)	Launched	Q2'14	61	300 W	$4235.00
Intel® Xeon Phi™ Coprocessor 7120D (16GB, 1.238 GHz, 61 core)	Launched	Q1'14	61	270 W	$4235.00
Intel® Xeon Phi™ Coprocessor 7120P (16GB, 1.238 GHz, 61 core)	Launched	Q2'13	61	300 W	$4129.00
Intel® Xeon Phi™ Coprocessor 7120X (16GB, 1.238 GHz, 61 core)	Launched	Q2'13	61	300 W	$4129.00

Intel Xeon Phi 31S1P 8GB

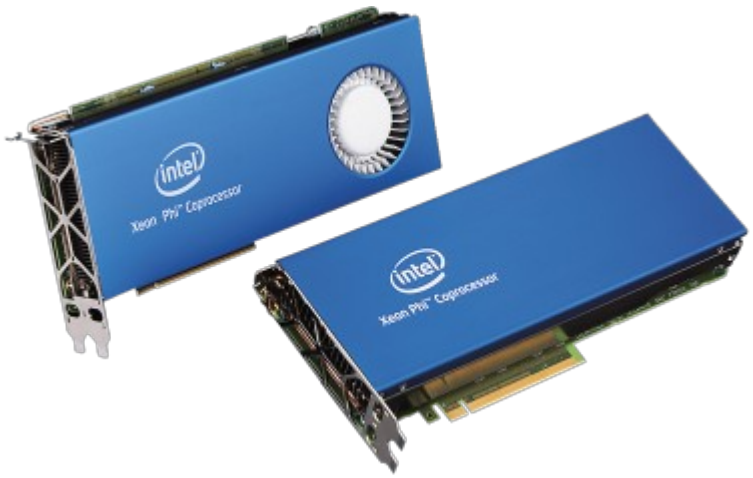
Listed Price: ~~$185.00~~

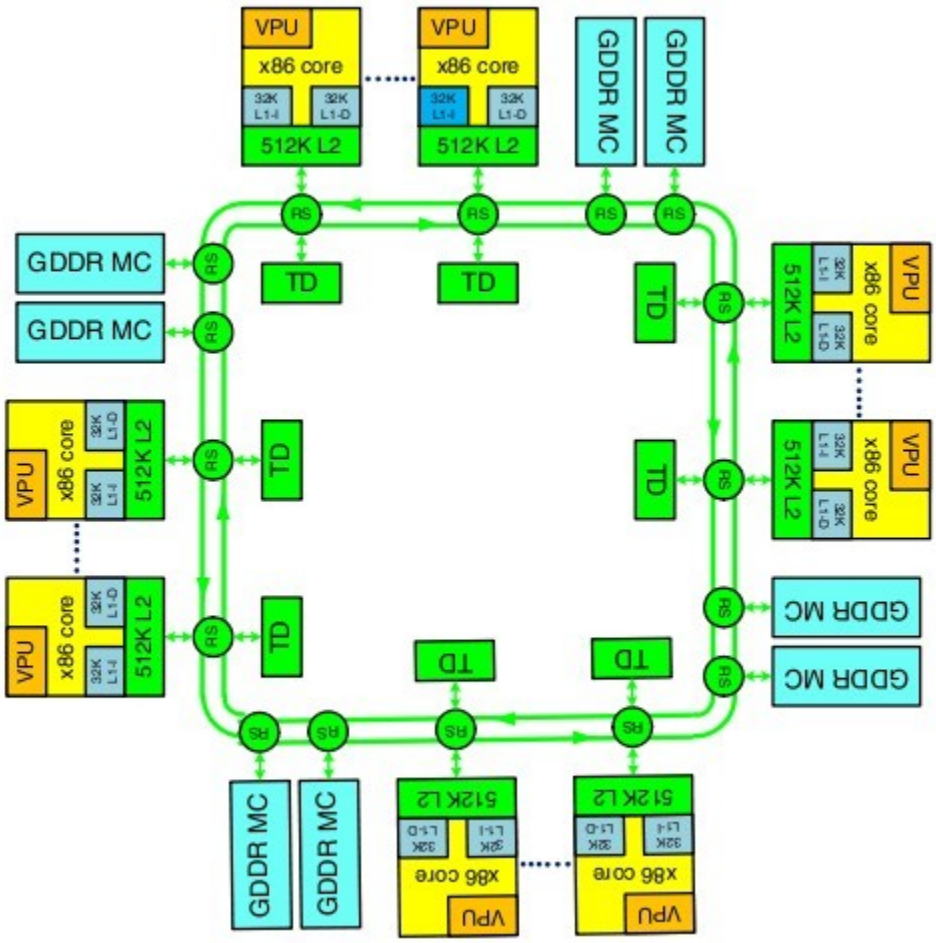
$150

with CODE IXPROMO

BUY NOW

Coupon code valid through 11/25/14.

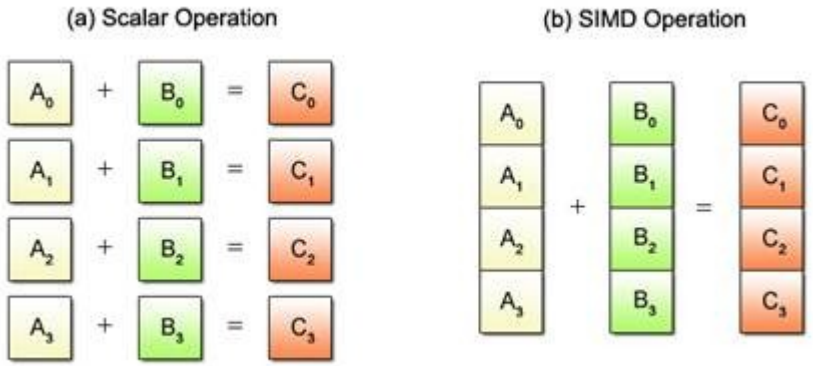



VPU
x86 core
32K L1-I
32K L1-D
512K L2
GDDR MC
RS
TD

- ilość rdzeni od 57 do 61 – zmodyfikowany P54C
- standardowy model programowania
- rdzenie x86 mogą uruchamiać standardowe instrukcje, np. EMT64T, **ale nie MMX, SSE czy AVX**
- jednostka wektora (512b) - SIMD
- 4 wątki sprzętowe
- podczas każdego cyklu uruchamiana są dwie instrukcje z każdego wątku (PipeU, PipeV)

Scalar:

```
int a[4] = { 1, 3, 5, 7 };
int b[4] = { 2, 4, 6, 8 };
int c[4];
c[0] = a[0] + b[0];     // 1 + 2
c[1] = a[1] + b[1];     // 3 + 4
c[2] = a[2] + b[2];     // 5 + 6
c[3] = c[3] + c[3];     // 7 + 8
```



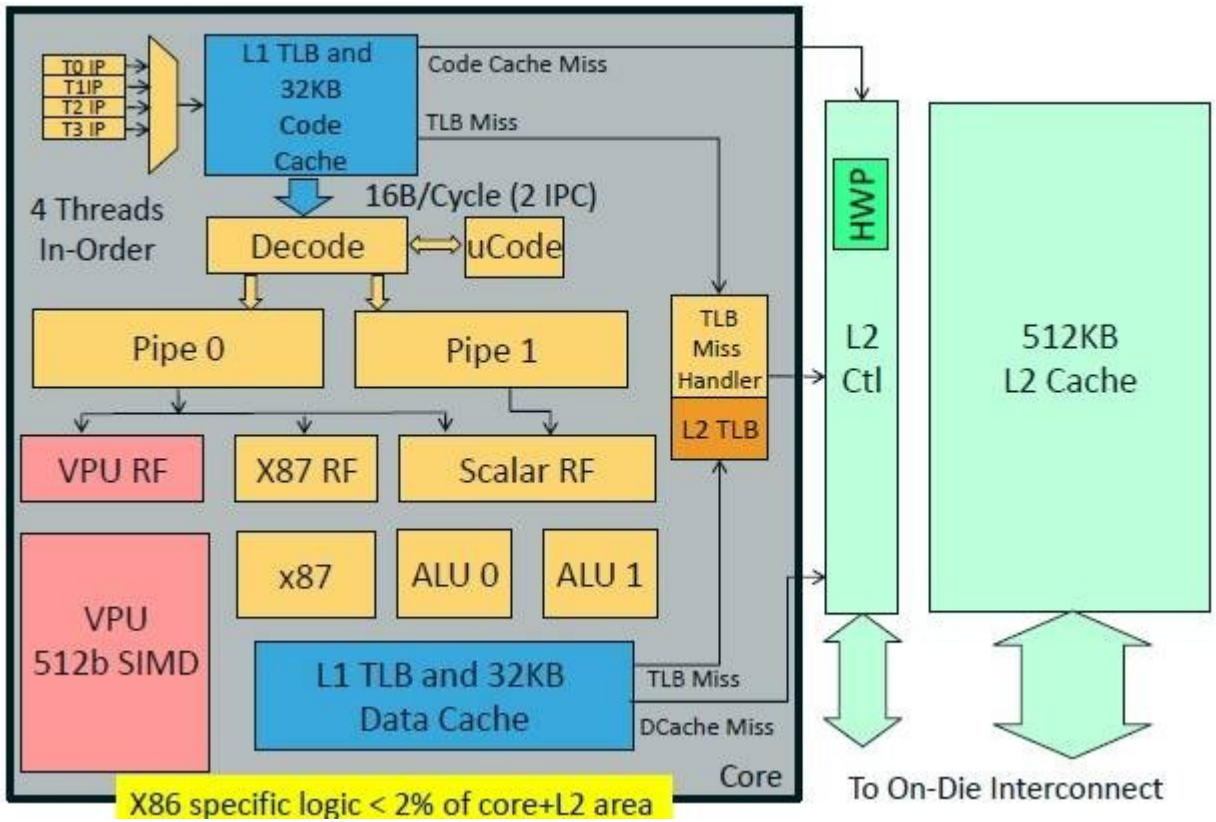
SIMD:

```
int a[4] __attribute__((aligned(16))) = { 1, 3, 5, 7 };
int b[4] __attribute__((aligned(16))) = { 2, 4, 6, 8 };
int c[4] __attribute__((aligned(16)));

__vector signed int *va = (__vector signed int *) a;
__vector signed int *vb = (__vector signed int *) b;
__vector signed int *vc = (__vector signed int *) c;

*vc = vec_add(*va, *vb);    // 1 + 2, 3 + 4, 5 + 6, 7 + 8
```

Źródło: http://bit.ly/1MbmiYx


T0 IP
T1IP
T2 IP
T3 IP
L1 TLB and 32KB Code Cache
Code Cache Miss
TLB Miss
4 Threads In-Order
16B/Cycle (2 IPC)
Decode
uCode
Pipe 0
Pipe 1
TLB Miss Handler
L2 TLB
VPU RF
X87 RF
Scalar RF
VPU 512b SIMD
x87
ALU 0
ALU 1
L1 TLB and 32KB Data Cache
TLB Miss
DCache Miss
Core
HWP
L2 Ctl
512KB L2 Cache
To On-Die Interconnect
X86 specific logic < 2% of core+L2 area

Wymagania dla HOST'a:

- PCIe 2.0 x16
- Base Address Registers **(BAR)** - PCIe potrzebuje adresować 64 bity

W BIOS'ie należy szukać:

- Memory Mapped I/O above 4GB
- PCI 64bit Resource Handling Above 4G Decoding
- MMIO above 4G
- Large BAR support

- Zasilacz ze złączami zasilania: 8 pin + 6 pin PCIe
- Turbina chłodząca

Więcej:

http://bit.ly/1kq2XrS

http://bit.ly/1H4infQ

Konfiguracje Hosta:

Procesor:2x Intel Xeon E5-2620 v2 (cena: 3800 PLN)

Płyta główna: Asus Z9PE-D8 WS (cena: 2106,26 PLN) **– EEB**

(Support Intel® Xeon Phi™ 3100 series (active fan SKU only) and NVIDIA Tesla K20C cards.)

Pamięc: 2x Kingston DDR3 8GB 1333MHz, 9CL ECC LV (cena: 584 PLN)

Zasilacz: Zasilacz Corsair RM 850W, 80PLUS GOLD, modularny, ATX (cena: 645 PLN)

Obudowa: Chenbro TW silver10566 EEB (cena: 384,01 PLN)

Suma: 7519 PLN

Minimalna:

Procesor: Intel CORE i5-4460 3.20GHz LGA1150 BOX (cena: 809 PLN)

Płyta główna: Asus GRYPHON Z97 ARMOR EDITION – **wymagany upgrade BIOS na wersje min. 2401** (cena: 733,52 PLN)

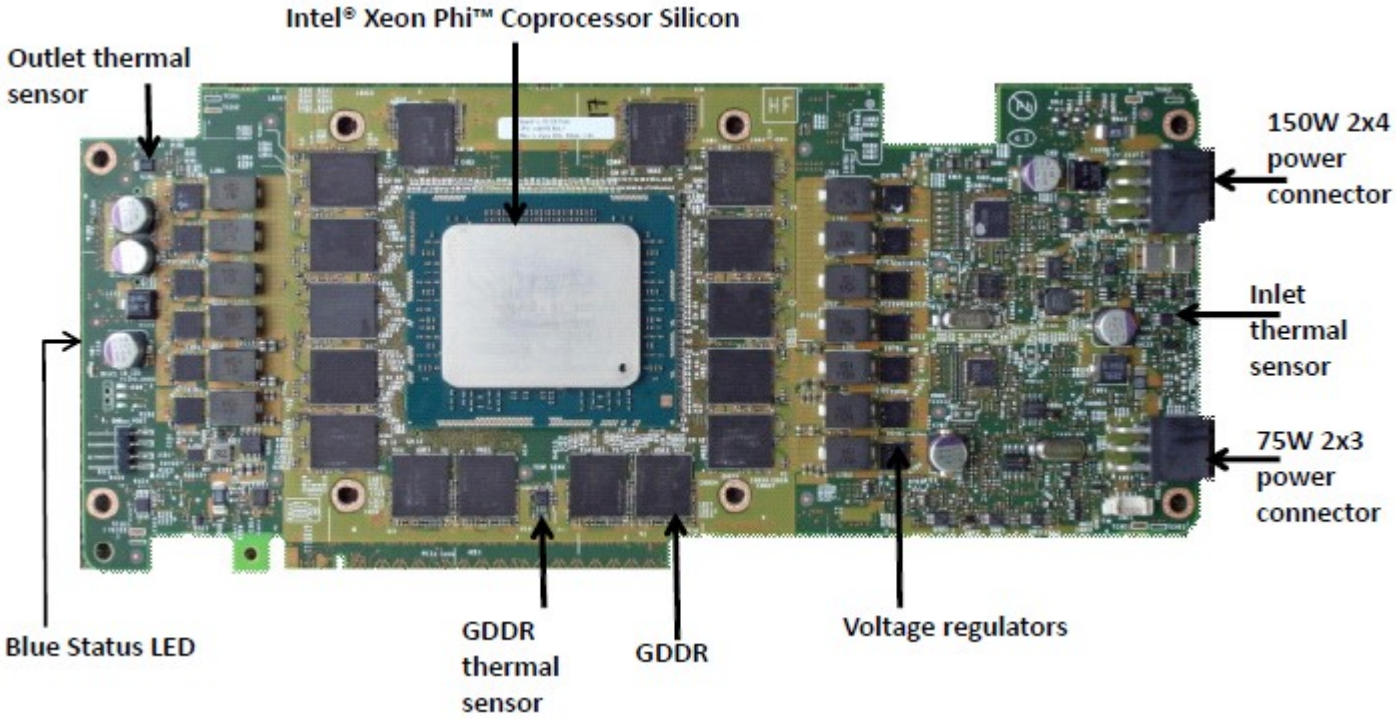
Pamięć: 2x Corsair Corsair Vengeance 4GB, DIMM,1600MHz, DDR3, CL9, XMP,Non-ECC 318 (cena: 318PLN)

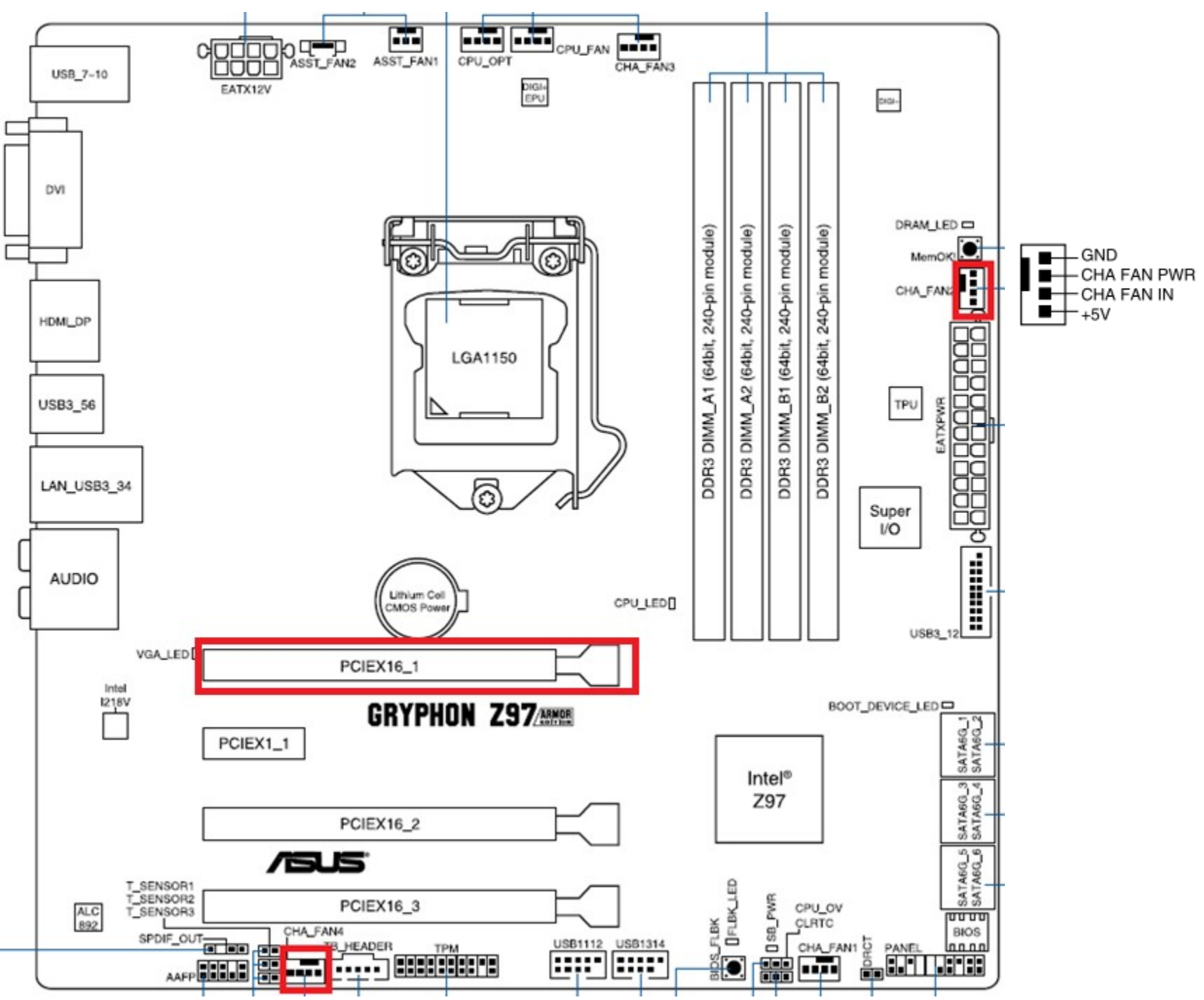
Zasilacz: Corsair CS Series 650W Modular 80+ GOLD (cena: 399PLN)

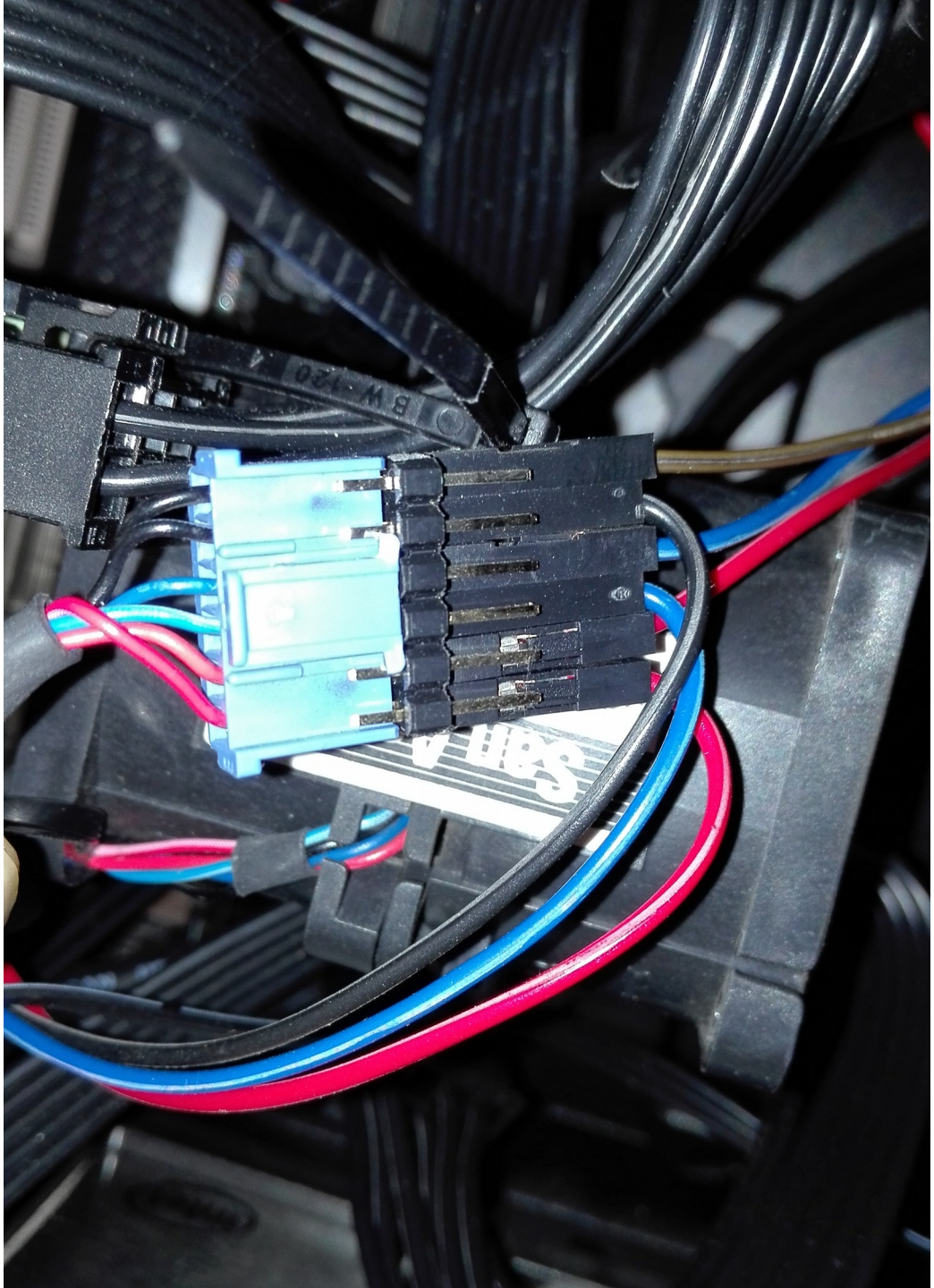
Obudowa: Cooler Master N200 NSE-200-KKN1 (cena: 145PLN)

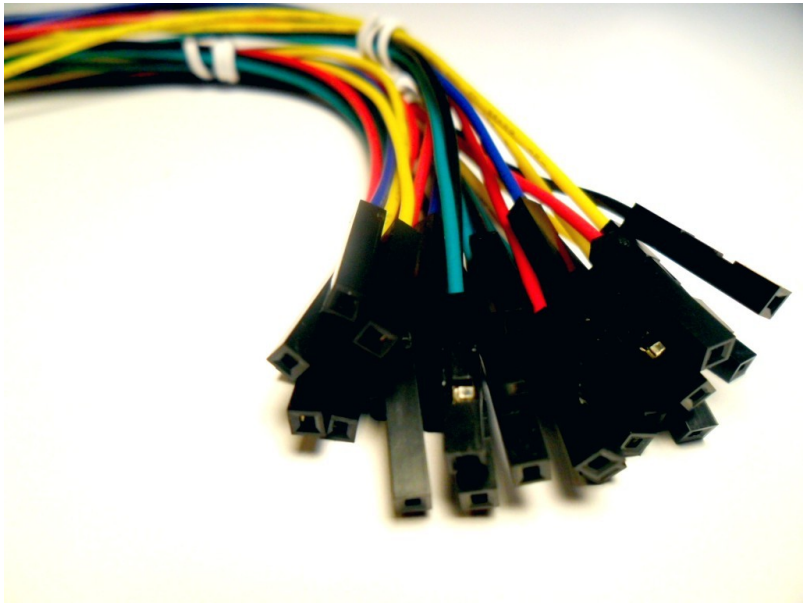
Suma: 2404,52 PLN

. Lead Wire ..... Inlet red ⊖black (Sensor) yellow (Control) brown
Outlet orange ⊖gray (Sensor) purple (Control) white




USB_7~10
EATX12V
ASST_FAN2
ASST_FAN1
CPU_OPT
CPU_FAN
CHA_FAN3
DIGI+ EPU
DIGI+
DVI
HDMI_DP
USB3_56
LAN_USB3_34
AUDIO
LGA1150
DDR3 DIMM_A1 (64bit, 240-pin module)
DDR3 DIMM_A2 (64bit, 240-pin module)
DDR3 DIMM_B1 (64bit, 240-pin module)
DDR3 DIMM_B2 (64bit, 240-pin module)
DRAM_LED
MemOK!
CHA_FAN2
GND
CHA FAN PWR
CHA FAN IN
+5V
TPU
EATXPWR
Super I/O
USB3_12
Lithium Cell CMOS Power
CPU_LED
VGA_LED
PCIEX16_1
Intel I218V
GRYPHON Z97 ARMOR edition
BOOT_DEVICE_LED
PCIEX1_1
Intel® Z97
SATA6G_1
SATA6G_2
SATA6G_3
SATA6G_4
SATA6G_5
SATA6G_6
PCIEX16_2
ASUS
T_SENSOR1
T_SENSOR2
T_SENSOR3
PCIEX16_3
ALC 892
SPDIF_OUT
CHA_FAN4
TB_HEADER
TPM
USB1112
USB1314
BIOS_FLBK
FLBK_LED
SB_PWR
CPU_OV
CLRTC
CHA_FAN1
DRCT
PANEL
BIOS
AAFP





San Ace 40
9CR0412S5038
DC 12V 1.1 A
SANYO DENKI
Made in China
EP 020624C R

P5: N/A

P6: N/A

X.M.P Disabled

Intel Rapid Storage Technology

On Off

FAN Profile

CPU FAN
1315 RPM

CHA1 FAN
8083 RPM

CHA2 FAN
12980 RPM

CHA3 FAN
N/A

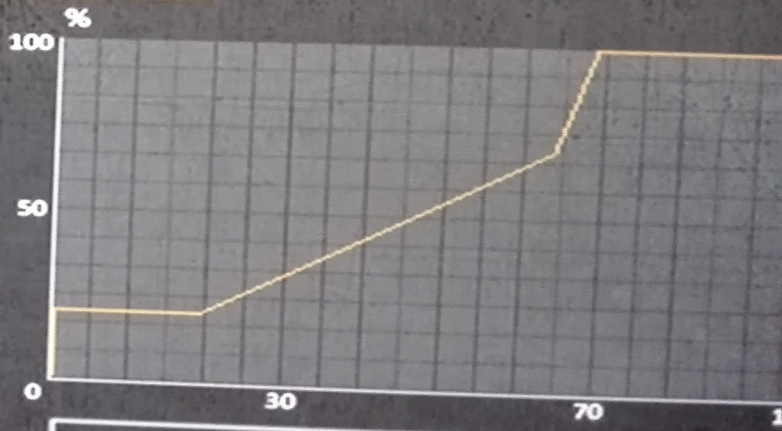
CHA4 FAN
N/A

CPU OPT FAN
N/A

ASST1 FAN
N/A

ASST2 FAN
N/A

CPU FAN



Manual Fan Tuning

Host:

Kernel do koprocesora,

uOS,

Konfiguracja: /etc/mpss/

Koprocesor:

pamięć flash,

SMC (System Management and Configuration) memory:

uruchamia firmware, zarządza konfiguracją oraz monitoruje koprocesor

1) ładowany mały osadzony kernel z flash

2) ładowany microkernel z hosta

3) ładowany uOS z HOST'a jako RAM FS

4) uOS w celu minimalizacji korzysta z busybox'a

Host:

OS: Centos 7.1

Kernel: **3.10.0-229.14.1.el7.x86_64**

- **Intel® Manycore Platform Software Stack (Intel® MPSS)** - http://intel.ly/1Rm2FNq
  (October 2, 2015 Intel® MPSS 3.6 released for Linux and Windows)
- Linux (mpss-3.6-linux.tar) for RedHat 6.5, RedHat 6.6, RedHat 6.7, RedHat 7.0, **RedHat 7.1**, SuSE SLES11 SP3, SuSE SLES11 SP4, SuSE SLES12.

MPSS wersja 3.6:

mpss-modules-2.6.32-431.el6.x86_64-3.6-1.x86_64.rpm
mpss-modules-2.6.32-504.el6.x86_64-3.6-1.x86_64.rpm
mpss-modules-2.6.32-573.el6.x86_64-3.6-1.x86_64.rpm
mpss-modules-3.0.101-63-default-3.6-1.x86_64.rpm
mpss-modules-3.0.76-0.11-default-3.6-1.x86_64.rpm
mpss-modules-3.10.0-123.el7.x86_64-3.6-1.x86_64.rpm
mpss-modules-3.10.0-229.el7.x86_64-3.6-1.x86_64.rpm

#cd /usr/src/mpss-3.6/

#rpm -ivh *.rpm

#cd /usr/src/mpss-3.6/src

#rpmbuild --rebuild mpss-modules-3.6-1.src.rpm

#cd /root/rpmbuild/RPMS/x86_64/

#rpm -ivh *.rpm

#dmesg

[0.150343] pci 0000:01:00.0: disabling BAR 0: [mem 0x00000000-0x1fffffff 64bit pref] (**bad alignment 0x200000000**)

[0.150351] pci 0000:01:00.0: BAR 4: assigned [mem **0xbf200000**-0xbf21ffff 64bit]

01:00.0 Co-processor: Intel Corporation Xeon Phi coprocessor 31S1 (rev 11)

Subsystem: Intel Corporation Device 2500

Flags: bus master, fast devsel, latency 0, IRQ 16

Memory at <unassigned> (64-bit, prefetchable) [size=8G]

Memory at **bf200000** (64-bit, non-prefetchable) [size=128K**]**

Capabilities: [44] Power Management version 3

Capabilities: [4c] Express Endpoint, MSI 00

Capabilities: [88] MSI: Enable- Count=1/16 Maskable- 64bit+

Capabilities: [98] MSI-X: Enable- Count=16 Masked-

Capabilities: [100] Advanced Error Reporting

Kernel driver in use: mic

https://github.com/xdsopl/mpss-modules/

mpss-modules-3.3.0 tested and working on linux-3.16.1

mpss-modules-3.4.3 tested and working on linux-3.19

mpss-modules-3.5.1 tested and working on linux-4.1

Kernel 3.19.8

#cp /boot/config-3.10.0-229.14.1.el7.x86_64 /usr/src/kernels/linux-3.19.8/.config

#cd /usr/src/kernels/linux-3.18.9/

#make olddefconfig

#make

#make modules_install

#make install

```
# dmesg | grep BAR

[    0.142446] pci 0000:01:00.0: BAR 0: assigned [mem 0x800000000-0x9fffffff 64bit pref]
[    0.142452] pci 0000:01:00.0: BAR 4: assigned [mem 0xbf200000-0xbf21ffff 64bit]



#lspci –v
01:00.0 Co-processor: Intel Corporation Xeon Phi coprocessor 31S1 (rev 11)
        Subsystem: Intel Corporation Device 2500
        Flags: bus master, fast devsel, latency 0, IRQ 16
        Memory at 800000000 (64-bit, prefetchable) [size=8G]
        Memory at bf200000 (64-bit, non-prefetchable) [size=128K]
        Capabilities: [44] Power Management version 3
        Capabilities: [4c] Express Endpoint, MSI 00
        Capabilities: [88] MSI: Enable- Count=1/16 Maskable- 64bit+
        Capabilities: [98] MSI-X: Enable+ Count=16 Masked-
        Capabilities: [100] Advanced Error Reporting
        Kernel driver in use: mic

#lsmod  | grep mic
mic                682460  12
```

- micflash
- micctrl
- micinfo
- micsmc
- miccheck
- micnativeloadex

#micflash –getversion

mic0: Flash read started

mic0: Read done

mic0: Version: 2.1.02.0391

mic0: Transitioning to ready state

micflash -update

No image path specified - Searching: **/usr/share/mpss**/flash

mic0: Flash image: /usr/share/mpss/flash/EXT_HP2_B1_0390-02.rom.smc

mic0: Flash update started

mic0: Flash update done

mic0: SMC update started

mic0: SMC update done

mic0: Transitioning to ready state

Please restart host for flash changes to take effect

#micctrl --initdefaults

Tworzy pliki konfiguracji w /etc/mpss

```
# micctrl –config
mic0:
    Linux Kernel:   /usr/share/mpss/boot/bzImage-knightscorner
    Map File:       /usr/share/mpss/boot/System.map-knightscorner
    MPSSVersion:    3.x
    BootOnStart:    Enabled
    Shutdowntimeout: 300 seconds

    Root Device:   Dynamic Ram Filesystem /var/mpss/mic0.image.gz from:
        Base:      CPIO /usr/share/mpss/boot/initramfs-knightscorner.cpio.gz
        Overlay:   Filelist /var/mpss/sep /var/mpss/sep/sep3.15/sep.filelist on
        CommonDir: Directory /var/mpss/common
        Micdir:    Directory /var/mpss/mic0

    Network:       Static Pair
        Hostname:  ramirez-mic0
        MIC IP:    172.31.1.1
        Host IP:   172.31.1.254
        Net Bits:  24
        NetMask:   255.255.255.0
        MtuSize:   64512
        MIC MAC:   4c:79:ba:1c:1a:70
        Host MAC:  4c:79:ba:1c:1a:71
```

#miccheck

Executing default tests for host
  Test 0: Check number of devices the OS sees in the system ... pass
  Test 1: Check mic driver is loaded ... pass
  Test 2: Check number of devices driver sees in the system ... pass
  Test 3: Check mpssd daemon is running ... pass
Executing default tests for device: 0
  Test 4 (mic0): Check device is in online state and its postcode is FF ... pass
  Test 5 (mic0): Check ras daemon is available in device ... pass
  Test 6 (mic0): Check running flash version is correct ... pass
  Test 7 (mic0): Check running SMC firmware version is correct ... pass

Status: OK
micctrl -R
        mic0: shutting down
         mic0: ready
         mic0: booting /usr/share/mpss/boot/bzImage-knightscorner **← kernel ładowany z Host'a**
         mic0: online

lrwxrwxrwx 1 root root 38 Jun 14 21:26 **bzImage-knightscorner -> bzImage-2.6.38+mpss3.4.3-knightscorner**

http://people.seas.harvard.edu/~apw/stress/

./stress **--cpu 50** --timeout 10s

#micsmc -c

mic0 (cores):

Device Utilization: User: 26.54%, System: 1.92%, Idle: 71.54%

Per Core Utilization (57 cores in use)

Core #1: User: **25.28%,** System: 0.00%, Idle: 74.72%

Core #2: User: 25.28%, System: 0.00%, Idle: 74.72%

Core #3: User: 25.28%, System: 0.00%, Idle: 74.72%

Core #4: User: 25.28%, System: 0.00%, Idle: 74.72%

Core #5: User: 25.28%, System: 0.00%, Idle: 74.72%

Core #6: User: 25.28%, System: 0.00%, Idle: 74.72%

Core #7: User: 25.28%, System: 0.00%, Idle: 74.72%

Core #8: User: 25.28%, System: 15.28%, Idle: 59.44%

./stress **--cpu 70** --timeout 10s

mic0 (cores):
Device Utilization: User: 30.80%, System: 1.91%, Idle: 67.29%
Per Core Utilization (57 cores in use)
Core #1: User: 25.00%, System: 0.28%, Idle: 74.72%
Core #2: User: 50.00%, System: 0.00%, Idle: 50.00%
Core #3: User: 25.00%, System: 0.00%, Idle: 75.00%
Core #4: User: 25.00%, System: 0.00%, Idle: 75.00%

./stress **--cpu 227** --timeout 20s

Device Utilization: User: 96.15%, System: 1.72%, Idle: 2.13%
Per Core Utilization (57 cores in use)
Core #1: User: 99.72%, System: 0.28%, Idle: 0.00%
Core #2: User: 100.00%, System: 0.00%, Idle: 0.00%
Core #3: User: 100.00%, System: 0.00%, Idle: 0.00%
Core #4: User: 100.00%, System: 0.00%, Idle: 0.00%
Core #5: User: 100.00%, System: 0.00%, Idle: 0.00%
…

mic0 (temp):

Cpu Temp: ..... 38.00 C

Memory Temp: ..... 30.00 C

Fan-In Temp: ..... 22.00 C

Fan-Out Temp: ..... 30.00 C

Core Rail Temp: ..... 27.00 C

Uncore Rail Temp: ..... 28.00 C

Memory Rail Temp: ..... 28.00 C

#uptime

19:38:59 up 35 min, 1 user, load average: **0.27**, 0.36, 0.30

./stress **--cpu 227** --timeout 600s

mic0 (temp):

Cpu Temp: ..... 58.00 C

Memory Temp: ..... 42.00 C

Fan-In Temp: ..... 24.00 C

Fan-Out Temp: ..... 44.00 C

Core Rail Temp: ..... 35.00 C

Uncore Rail Temp: ..... 36.00 C

Memory Rail Temp: ..... 36.00 C

uptime

19:45:07 up 41 min, 1 user, load average: **228.70**, 146.29, 72.56

- Offload

```
#include <stdio.h>
#include <unistd.h>
int main(int argc, char *argv[]){
        printf("cpus: %d \n", sysconf(_SC_NPROCESSORS_ONLN )); fflush(0);
#pragma offload target(mic)
    {   printf("cpus: %d \n", sysconf(_SC_NPROCESSORS_ONLN )); fflush(0); }
}
# gcc offload.c -o offload
[root@ramirez src]# ./offload
cpus: 4
cpus: 4

[root@ramirez src]# icc offload.c -o offload
[root@ramirez src]# ./offload
cpus: 4
cpus: 228
```

#iptables -I INPUT -s 172.31.0.0/16 -j ACCEPT
#cat /etc/exports
/mic0fs mic0(rw,no_root_squash)

#exportfs -a
#showmount -e 172.31.1.254
Export list for 172.31.1.254:
/mic0fs ramirez-mic0.hb.pl

--- mic0
#mount -o nolock -t nfs 172.31.1.254:/mic0fs /mnt/host/

[root@ramirez-mic0 ~]# df -h
...
172.31.1.254:/mic0fs 27.4G 2.7G 23.3G 10% /mnt/host

[root@ramirez ~]# **micctrl --addnfs=/mic0fs --dir=/mnt/host**

Plik (/var/mpss/mic0/etc/fstab)
cat /var/mpss/mic0/etc/fstab
...
172.31.1.254:/mic0fs /mnt/host nfs defaults 1 1

Overlay

/var/mpss/pstree/

#ls

pstree pstree.filelist

#cat pstree.filelist

dir /usr/bin 755 0 0

file /usr/bin/pstree pstree/pstree 755 0 0

#ls /etc/mpss/conf.d/

_amplxe_vtune_amplifier_xe_2016.1.0.424694.conf pstree.conf

_amplxe_vtune_amplifier_xe_2016.1.0.424694_itt.conf sep.conf

#cat pstree.conf

Overlay Filelist /var/mpss/pstree /var/mpss/pstree/pstree.filelist on

#micctl -R

micctrl –config

…

Overlay: Filelist /var/mpss/pstree /var/mpss/pstree/pstree.filelist on

…

Język R dla PHI:

Revolution R Open
https://mran.revolutionanalytics.com/ - (RHEL / CentOS 7 - RRO MKL)
https://software.intel.com/en-us/intel-mkl

#export
MIC_LD_LIBRARY_PATH=/opt/intel/compilers_and_libraries_2016.0.109/linux/mkl/lib/intel64_lin_mic:/usr/lib/compiler/lib/mic/
export MKL_MIC_ENABLE=1
export MKL_MIC_DISABLE_HOST_FALLBACK=1

#Rscript test.R

AO library failed to initialize! - gdy brak MKL z INTEL'a !

Więcej: http://bit.ly/1GmH6JZ

Dokument do pobrania na:

http://xeonphi.pl