PDC Center for High Performance Computing

Michaela Barth caela@kth.se
Gert Svensson gert@kth.se

Dardel sustainability at a glance

Dardel

"In Sweden, KTH's new Dardel system serves as a lovely new canvas for existing artwork by Nils Dardel, depicting a story by author Thora Dardel (his wife) as well as his portrait of her."
(honourable mention in HPCwire 2023 Superlative Awards in the category 'Favorite Cabinet Art')

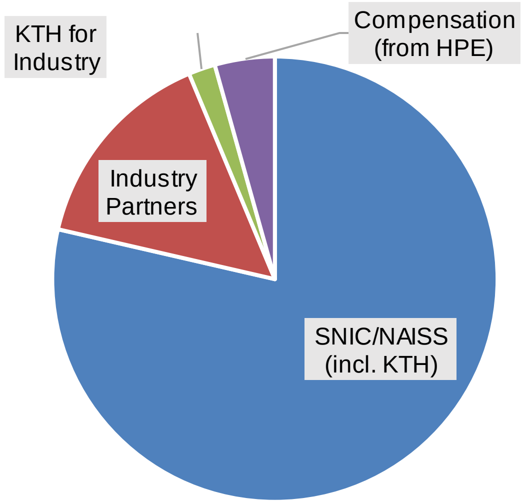
Expansion Autumn 2022

KTH VETENSKAP OCH KONST
PDC
Dardel
Dardel

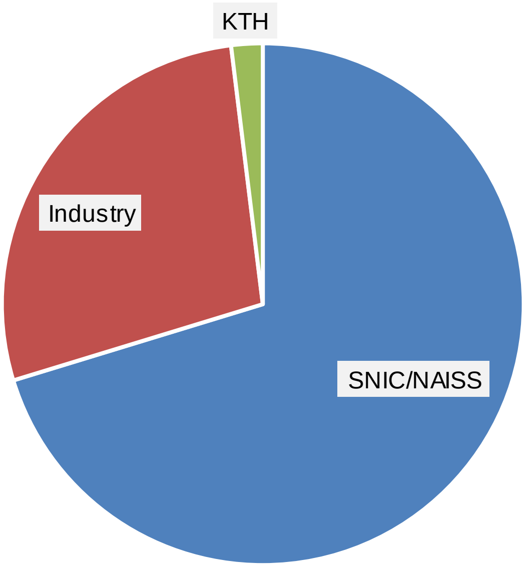
Financing a supercomputer

Initial cost share SNIC/NAISS including extension and all addendums: ~142 MSEK

NAISS National Academic Infrastructure for Supercomputing in Sweden



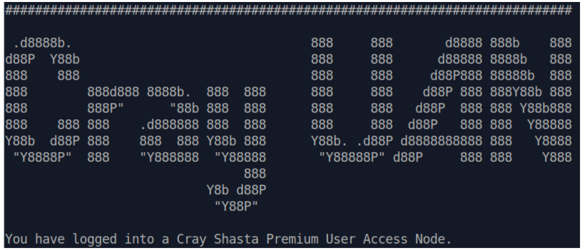
Inititial cost (hardware and install))



Variable costs for 5 years

Building blocks

- Dardel CPU part
- Dardel GPU part (since extension)
- Storage
- "Adminstrative": Management nodes, login nodes, file transfer nodes, LDAP (+ login portal)
- Interconnect
- Power
- Cooling



```
############################################################################

 .d8888b.                                888     888       d8888 888b    888
d88P  Y88b                               888     888      d88888 8888b   888
888    888                               888     888     d88P888 88888b  888
888        888d888 8888b.  888  888      888     888    d88P 888 888Y88b 888
888        888P"      "88b 888  888      888     888   d88P  888 888 Y88b888
888    888 888    .d888888 888  888      888     888  d88P   888 888  Y88888
Y88b  d88P 888    888  888 Y88b 888      Y88b. .d88P d8888888888 888   Y8888
 "Y8888P"  888    "Y888888  "Y88888       "Y88888P" d88P     888 888    Y888
                                888
                           Y8b d88P
                            "Y88P"

You have logged into a Cray Shasta Premium User Access Node.
```

Lustre Storage

2 totally separate **HPE ClusterStor E1000** Lustre systems "klemming" and "scania"
Estimated power released into air for klemming disk system: **32 kW**

Klemming:

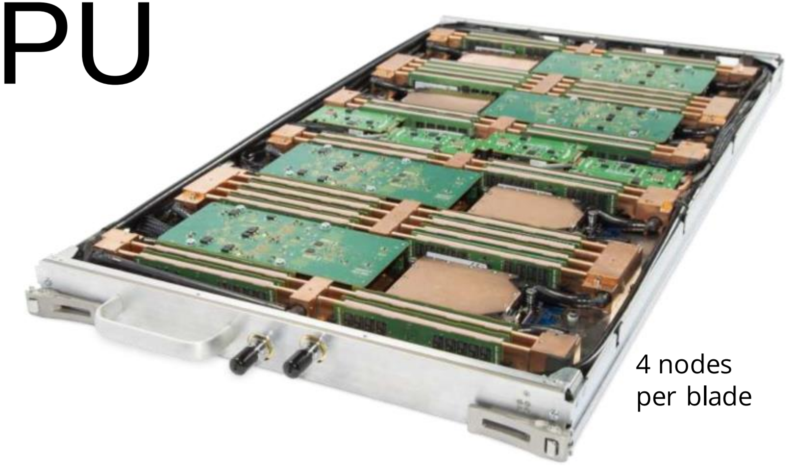
- **12 PB** user space
- Raw Capacity > 24 PB
- Peak Performance: **180 GB/s**
  (streaming speed for large files; IOR benchmark)

Storage Servers and Modules:

- 1 System Management Unit (SMU)
- 2 Metadata Server Units (MDU) hosting two meta data servers each
- 12 Scalable Storage Unit SSU-D2 Servers with two 4U106 Disk Storage Modules each

2 Object Storage Targets per 4U106; 2 HDD JBODs and 4 OSTs per SSU-D2 ⇒ **48 OSTs**
SSUs, SMU and MDUs have more or less the same hardware

Dardel-CPU



4 nodes per blade

HPE CRAY EX system à 1270 nodes

Different memory sizes (altogether **> 160 PB**):

- 736 × 256 GB "thin"
- 304 × 512 GB "large"
- 8 × 1024 GB "huge"
- 10 × 2048 GB "giant"

Mean power consumption CPU part for 554 nodes (without storage):
357 kW water-cooled, 25 kW air-cooled, observed **peaks at 378 kW**

Mean power consumption estimated as the time average for running each 50% of the time:

- Gromacs throughput
- NEK5000 strong scaling

benchmarks

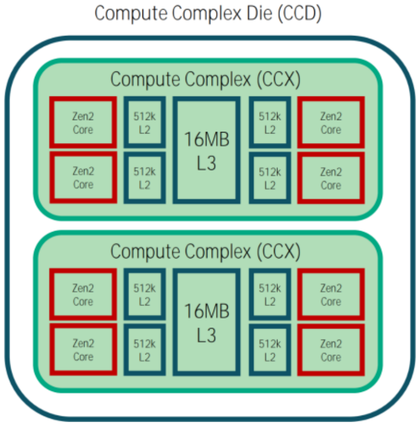
1 node:

Custom AMD EPYC^{TM} 7742 "Zen2 Rome" 2.25GHz with 128 cores

Dardel-CPU

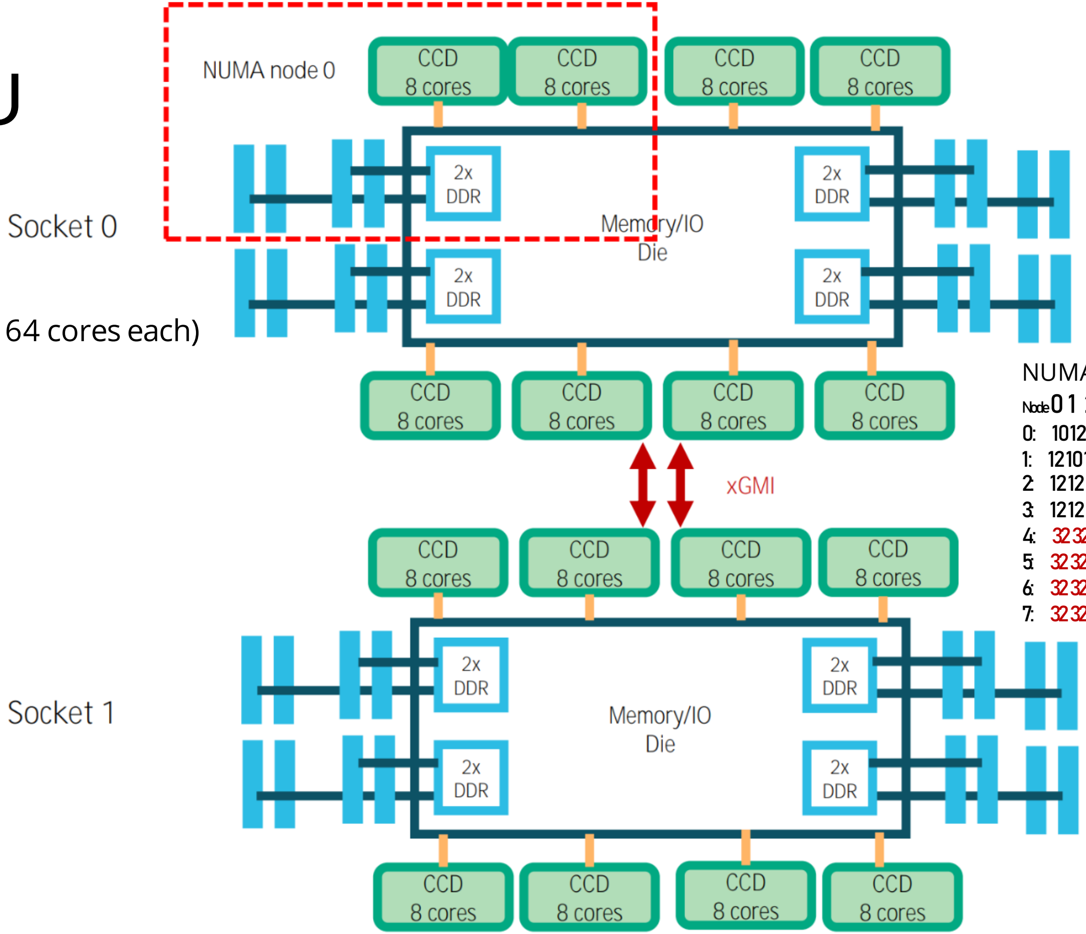
1 node:

AMD EPYC™ 7742 "Zen2 Rome"
(2 sockets per node with 2.25GHz and 64 cores each)
organised in 2×4 NUMA nodes



16 Compute Complex Dies (CCDs)

- 7 nm process
- Infinity Fabric™ interconnect
- host cores and L2/L3 cache

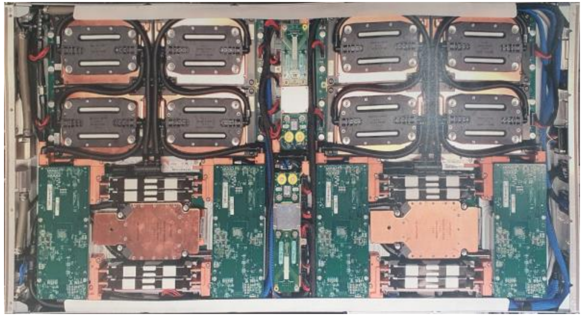


NUMA node distances:

Node	0	1	2	3	4	5	6	7
0:	10	12	12	12	32	32	32	32
1:	12	10	12	12	32	32	32	32
2:	12	12	10	12	32	32	32	32
3:	12	12	12	10	32	32	32	32
4:	32	32	32	32	10	12	12	12
5:	32	32	32	32	12	10	12	12
6:	32	32	32	32	12	12	10	12
7:	32	32	32	32	12	12	12	10

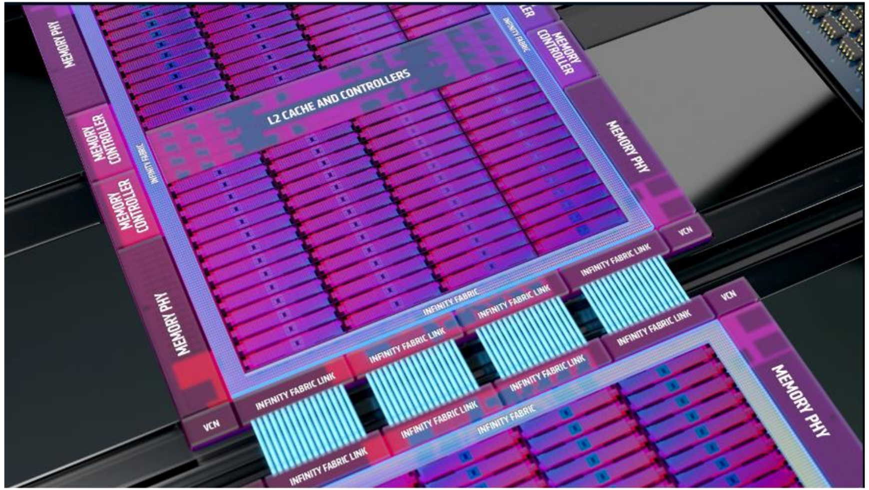
Dardel-GPU

28 blades with 2 nodes each



56 HPE Cray EX235a nodes

- AMD EPYC$^{\text{TM}}$ 7A53 "Trento" (special version) 64-core 2.32 GHz processor
- Four Instinct$^{\text{TM}}$ MI250X GPUs as accelerators
- 6 nm process
- **123.5 kW water-cooled, 8kW air-cooled**



Mean power consumption estimated as the time average for running each 50% of the time:

- Gromacs throughput
- PyFR double precision strong scaling benchmarks

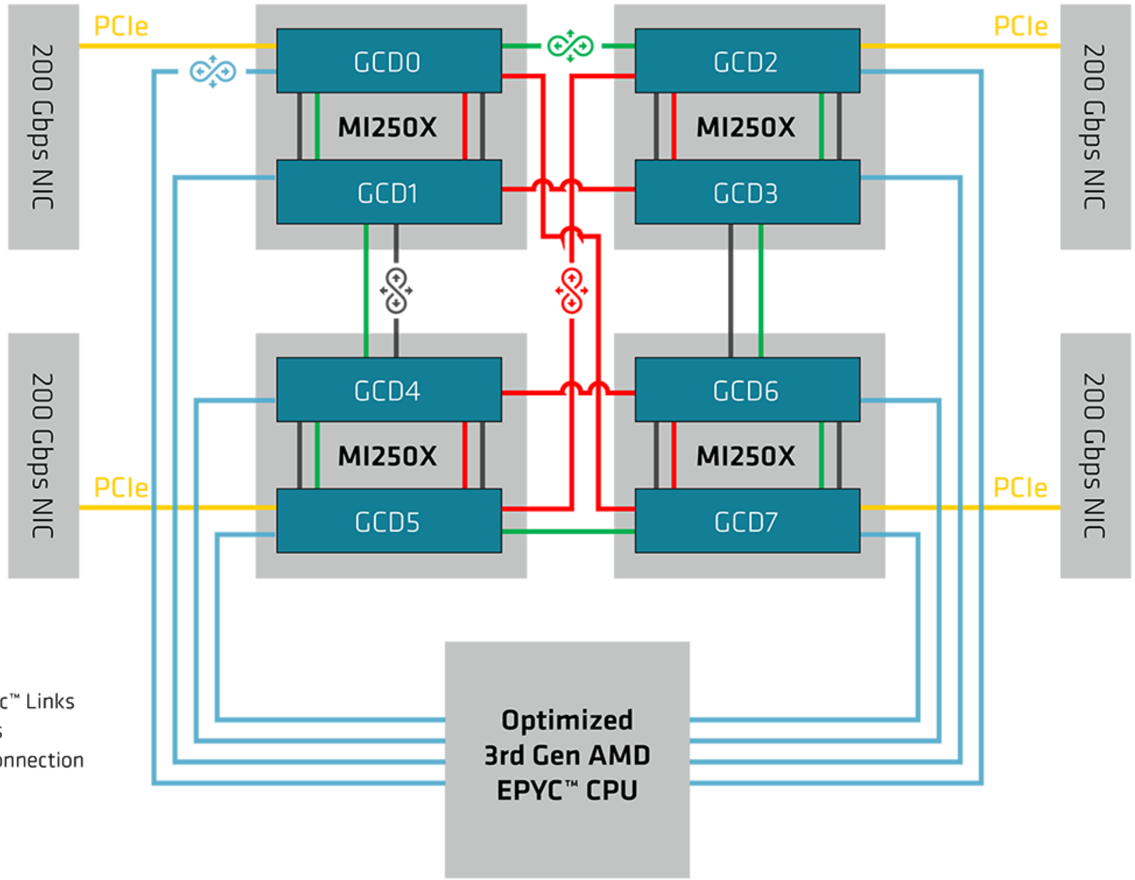
GPU node close up

Four AMD Instinct™ MI250X GPUs (performance of up to 95.7 TFLOPS in double precision) with two Graphics Compute Dies à 110 compute units ("cores") each
(⇒ software wise every node has 8 GCDs and a total of 880 compute units)

512 GB of shared fast HBM2E memory (64 GB for each die)
cache-coherent to simplify programming

Connected by AMD’s Infinity Fabric®

Optimized 3rd Gen AMD EPYC™ Processor + AMD Instinct™ MI250X Accelerator

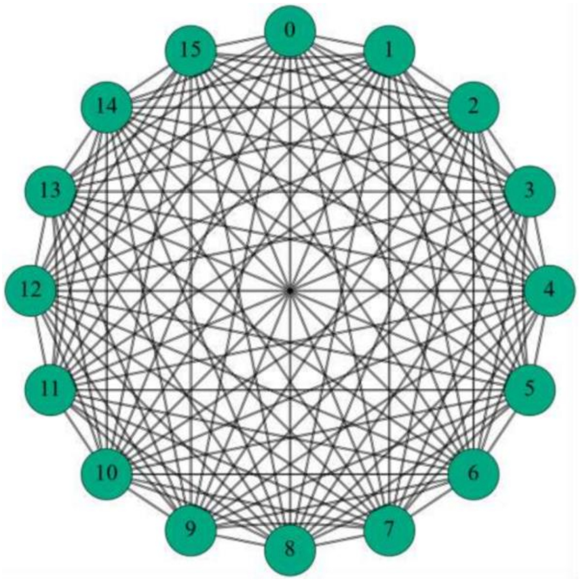


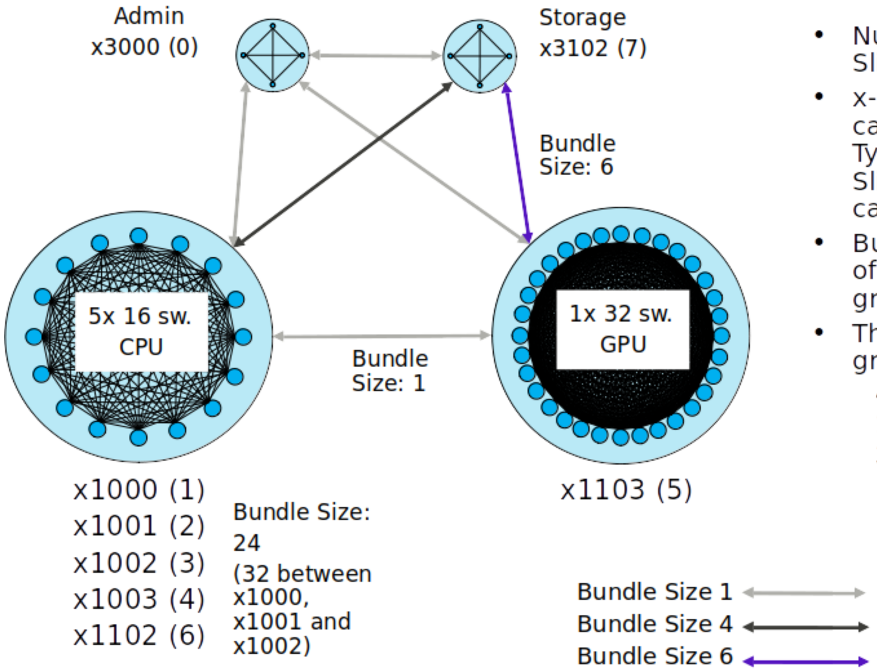
🟩 🟥 ⬛ 🟦 Green, Red, Gray, and Blue lines are AMD Infinity Fabric™ Links
Red and Green links can create two bi-directional rings
Blue Infinity Fabric Link provides coherent GCD-CPU connection

🟨 Orange lines are PCIe® Gen4 with ESM

High-Speed-Network

- Interconnect is HPE Slingshot (ethernet-based) using Dragonfly topology
- 200 Gb/s since March 2023
- Five CPU groups
- + 1 GPU group





Admin
x3000 (0)

Storage
x3102 (7)

Bundle
Size: 6

5x 16 sw.
CPU

Bundle
Size: 1

1x 32 sw.
GPU

x1000 (1)
x1001 (2)
x1002 (3)
x1003 (4)
x1102 (6)

Bundle Size:
24
(32 between
x1000,
x1001 and
x1002)

x1103 (5)

Bundle Size 1
Bundle Size 4
Bundle Size 6

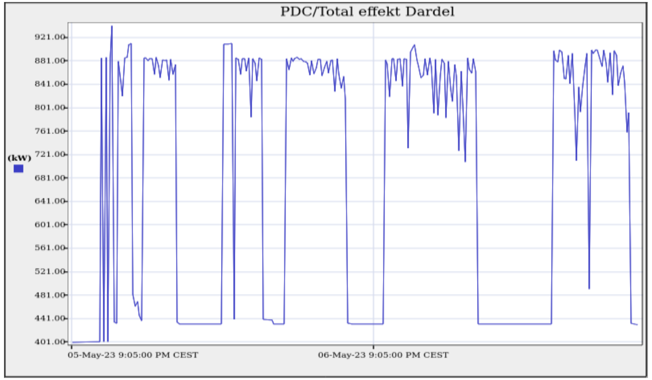
- Numbers in brackets are Slingshot group IDs.
- x-Numbers are names of cabinets and racks. Typically there is one Slingshot group per cabinet.
- Bundle size is the number of cables between groups/cabinets.
- Three types of Slingshot groups:
  - 4-sw(itch) – Admin, Storage
  - 16-sw(itch) – CPU nodes
  - 32-sw(itch) – GPU nodes

Worldwide ranking

Dardel-GPU:

- Fastest in Sweden
- **#5** on the **Green500**
- #77 in Top500, after entering as #**68** in 2022
- R_{max} **8.26 PFlop/s** (Maximal LINPACK performance achieved)

Nominal number of cores includes CPUs: 56*(64+880)=52,864, R_{peak} above 10.2 PFlop/s
Frontier (#1 Top 500) and LUMI (#3 Top 500) are #6 and #7 in Green 500



Dardel: CPU:

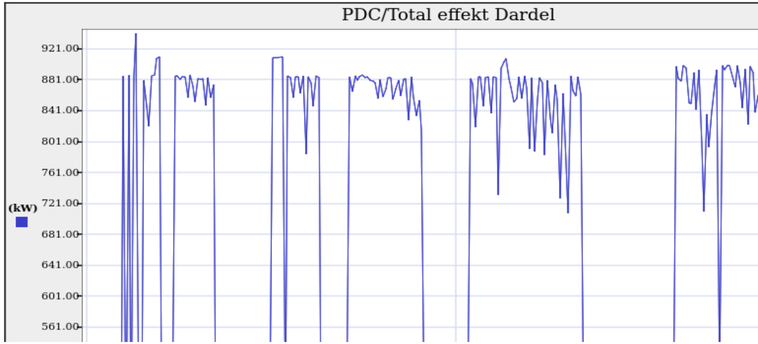
- #153 in Top500 after expansion, entered on #287 in 2021
- R_{max} **4.08 PFlop/s** running on 1024 (of 1270) nodes (≅131,072 CPU cores)

Energy efficiency

Comparison to previous system Beskow:
1.8 PF HPL, 156.4 TB memory total; 67,456 cores; typical 740 kW (~2.5 GFs p. W.)
Dardel CPU:
Before extension: 544 of 554 nodes (69632 cores, 141 TB): 2.28 PF HPL, 378 kW peak

- **CPU-only: ~6 GigaFlops per Watt**

After extension: 1024 (of 1270) nodes (131,072 cores)
4.08 PF HPL, 900 kW peak, (non-perfect configuration)



- **GPU part: ~60 GigaFlops per Watt**
  higher performance

Combine CPU and GPUs so power consumption compares to Beskow
⇒ a factor of 6 of performance increase
15-year-old trend that performance increases tremendously while power consumption almost stagnates

KTH BREAKS NEW GROUND IN SCIENCE USING AMD EPYC™ CPUS AND INSTINCT™ GPUS

Six times the compute performance in the same power envelope using AMD EPYC processors and AMD Instinct accelerators

AMD

EPYC | INSTINCT

Power balance

Total "mean" power consumption Dardel (all parts, after extension):
1065 kW ⇒ 9.3 GWh per year

- My own house needs 0.12% of that ⇒ Dardel burns my yearly usage in less than two days.
- Power envelope limitation: ~1.5 megawatt of power consumption.

Q: Is energy efficiency enough to reduce the environmental footprint?
A: No.

1. **Using energy-efficient hardware**

2. **Re-using produced heat**

3. **Use electricity from renewable resources**

Heat re-use history

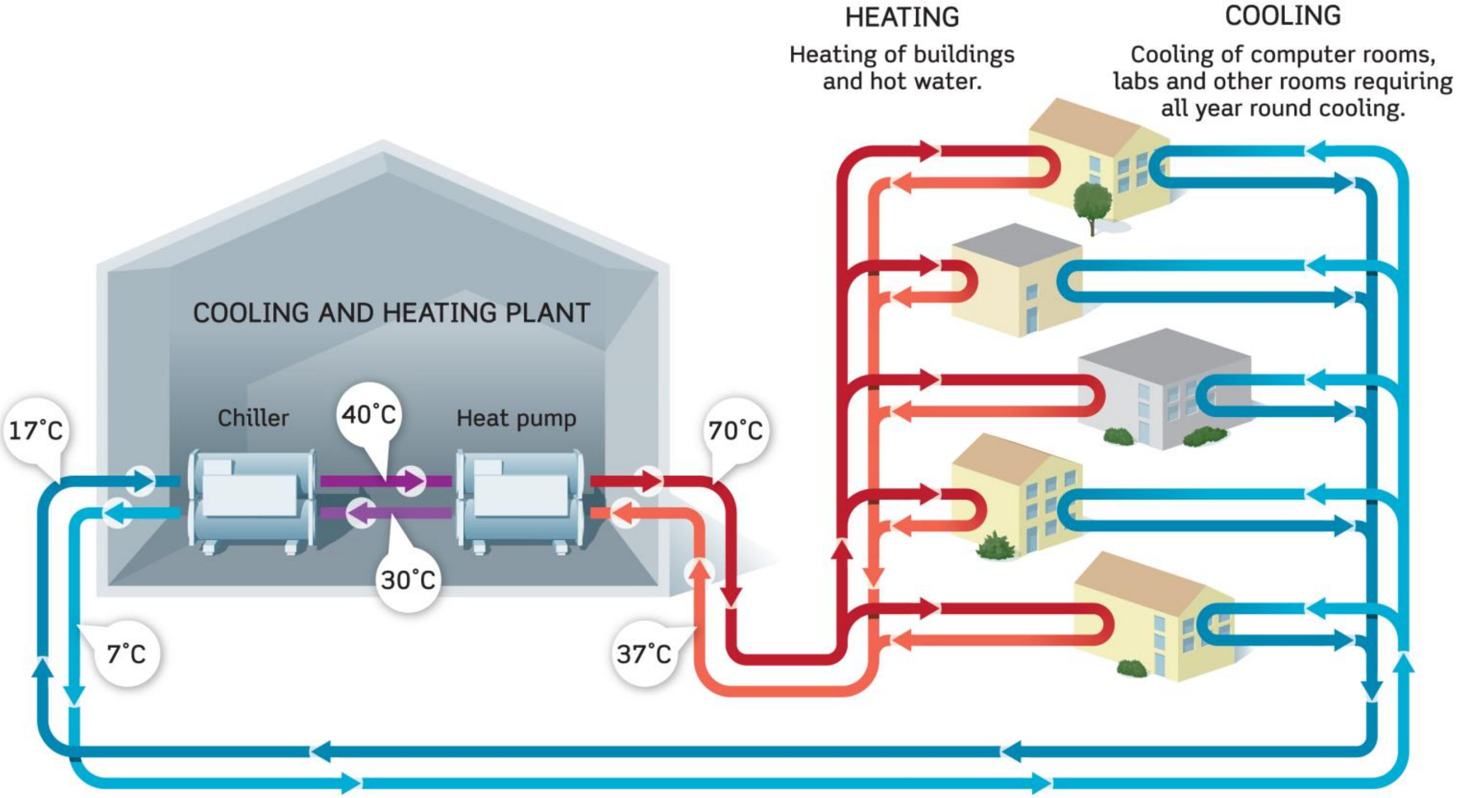
PDC as pioneer re-using heat from supercomputers since 2009:

Cray XE6 "Lindgren":
hot air ⇒ ventilator hoods ⇒ heat exchanger coils ⇒ water ⇒ Chemistry building

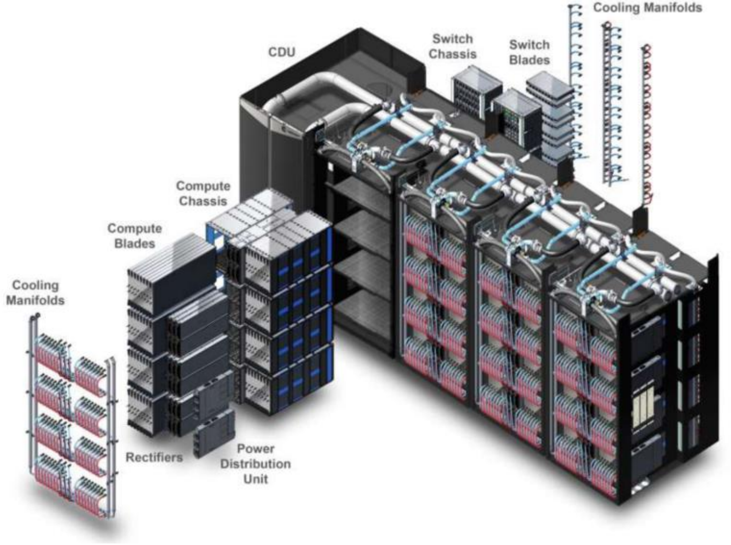
Cray XC40 "Beskow":
Air-to-water heat exchangers were included in the racks/cabinets
In 2015 new KTH main campus heat-pump installed

HPE Cray EX "Dardel":
Direct liquid-cooled
Connected to KTH heat pump via heat exchanger
Contributing to heat the buildings on KTH main campus

KTH heat re-use system


HEATING
Heating of buildings and hot water.
COOLING
Cooling of computer rooms, labs and other rooms requiring all year round cooling.
COOLING AND HEATING PLANT
Chiller
40˚C
Heat pump
17˚C
70˚C
30˚C
7˚C
37˚C

Cooling



Dardel CPU and GPU parts are **directly liquid-cooled** (storage and other surrounding infrastructure are not).

- One Cooling Distribution Unit (CDU) per row

A small percentage of produced heat still goes into the air.

At mean usage: **about 120 kW goes into the air, an additional 20% of power is needed to cool** that part.

To avoid overheating we must

- use even more electricity (~30%) to cool the system (€€€)

or

- **lead the heat where we can make good use of it.**

Both air and liquid are cooled down via heat exchangers and connected to the KTH heat pump, as part of a heating/cooling network.

Significant savings during winter for KTH when heating buildings!

Committed to sustainability

Continued commitment to operate our systems as environmentally friendly as possible:

100% of the electricity that PDC uses is **from renewable sources**, both for powering and cooling.

7 AFFORDABLE AND CLEAN ENERGY

KTH embraces the **UN Sustainable Development Goals**.

KTH's electricity supply agreement requires all the electricity that is provided to KTH to come from renewable sources.

Reporting is done yearly to the Swedish Environmental Protection Agency.

SUSTAINABLE DEVELOPMENT GOALS

KTH ROYAL INSTITUTE
OF TECHNOLOGY