GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Efficient HPC datacenter operation with direct liquid cooling

Experience with the NHR System Emmy

Sebastian Krey, Christian Boehme, Julian Kunkel



28.04.2022

Outline

1 HPC Systems in the MDC at GWDG/Uni Göttingen

2 Modular Data Center

3 First year of operation

4 Cooling optimization

5 Results

6 Summary and Outlook

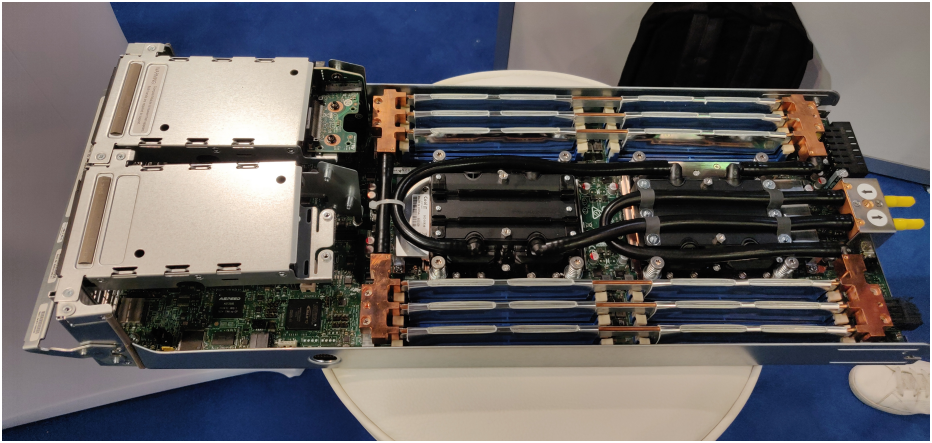
HLRN-IV “Emmy” and GWDG SCC

Uni Göttingen/GWDG

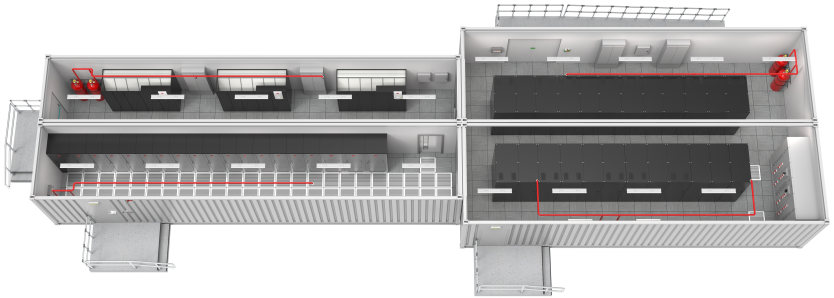
- TOP 500: #47 in 2020-11 (5.95 PFlop/s), now #82,
  approx (inofficial) 4.56 GFlops/Watt (would have been #55 in 2020-11)
- phase 1 compute nodes (air cooled)
  - 2x Intel Xeon Gold 6148 (SKL), 40 cores per node, 480 GB SSD
  - 432x 192 GB, 16x 768 GB
  - 19 GPU nodes: 2x Intel Xeon Gold, 4-8x Tesla V100 32 GB or 4x Quadro RTX 5000, 192-768 GB
- phase 2 compute nodes (warm water DLC)
  - 2x Intel Xeon Platinum 9242 (CLX-AP), 96 cores per node
  - 1100x 384 GB, 16x 768 GB, 2x 1536 GB
- file systems
  - WORK: 8.6 PiB DDN Lustre (130 TiB SSDs) + 130 TiB DDN IME Burst Buffer
  - HOME: 340 TiB DDN GPFS, PERM: GWDG tape archive (Quantum StorNext)
  - SCC Work: 2.1 PiB BeeGFS (100 TiB SSDs), GWDG Home (Quantum StorNext)

CLX-AP compute node with direct water-cooling

Intel S9200WK Walker Pass blade



Modular Data Center (MDC) from Stulz TI



- two rooms built from joint containers (air- and warm-water-cooled zone)
- space for **19 air-cooled and 14 warm-water-cooled racks**
- tailor-made for HLRN/SCC system needs

Modular Data Center (MDC)

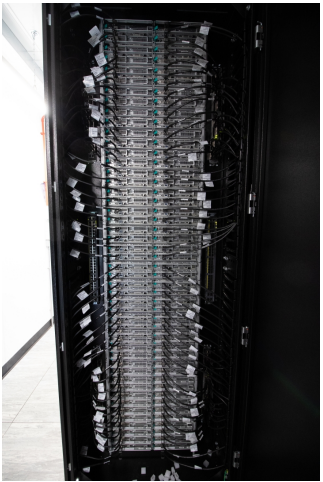
- used since July 2020, first for relocating HLRN P1 to allow combined operation of P1 and P2
- since August 2020: installation of HLRN P2, SCC Expansion
- full operation since October 2020
- good power efficiency (full datacenter PUE 1.10 at max load during acceptance testing, at 19.6°C average primary supply temperature)






Bull
atos technologies









MDC design targets

- capacity for 1.4 MW peak IT load
- space for 19 air cooled racks (11 compute, 5 infrastructure, 3 storage)
- air cooled capacity up to 300 kW
- space for 14 DLC racks (approx 78-80% heat capture rate)
- at least 1100 kW capacity for DLC systems with 32°C supply temperature (temporary increase up to 36°C possible)
- peak load operation at up to 19°C ambient, 75% at 32°C ambient
- full system PUE of at least 1.14

MDC cooling infrastructure

- 2 dry air coolers, each providing 660 kW cooling capacity at 32°C ambient for 36°C primary supply temperature
- 3 downflow CRAC units, each has 105 kW capacity, two compressor ciruits and a free cooling circuit allowing mixed mode operation
- 6 in-row coolers, each has 37.5 kW capacity, a speed controlled compressor and a free cooling circuit allowing mixed mode operation
- 3 CoolIT end of row CDUs to create 3 building blocks consisting of 4 compute racks, 2 in-row coolers and one CDU.

First year of operation

- Air cooled compute racks peak power consumption 23.5 kW per rack
- Warm water cooled racks peak power consumption 96 kW per rack
- Long term average power consumption in normal operation 80% of peak (70kW per rack)
- Typical full load power consumption 88-92% of peak (85-88kW per rack)
- Powerusers with highly optimized code constant 95-97% of peak (91-93kW per rack)
- Regular power swings of 400 kW (start of two 256 node jobs) within less than 3 mins (up to 120 kW for air cooled systems).
- Power swings up to 600 kW happen from time to time

First year of operation

- Full data center PUE 1.13 +0.07/- 0.03
  (includes cooling, UPS for storage and infrastructure, environmental monitoring, video surveillance, access control, fire extinguishing and alert systems, lights, cleaner sockets, 1.13=1.045x1.081)
- PUE for air cooled systems 1.24 + 0.15/-0.07
  (includes UPS, 1.24=1.045x1.187)
- PUE for DLC cooled systems 1.07 +0.02/- 0.03
  (DLC with in-row coolers for residual heat, 1.07=1.045x1.024)
- Average primary supply temperature 17.8°C, 33.4 hours above 32°C, 6.8 hours above 36°C (load reduction necessary)
- 98.7% of the year primary supply above 8°C, 70% above 14°C

Observations during operation and cooling experiments

Air cooled systems

- Large power consumption increase of air cooled systems at cold aisle temperatures >26°C (fans)
- Very little differences at lower temperatures
- PUE plateau at primary temperatures between 10°C and 14°C under high load (up to 15°C under light load)
- PUE improvement at primary temperatures <10°C very small (less than 0.03)
- At low loads 78% CRAC fan speed sufficient
- At high load cooling performance increases up to 87% CRAC fan speed
- Temperature oscillations due to power cycling of compressor circuits at the same temperatures

Research questions

- How exact are the CDU cooling duty estimates and the heat capture rate calculations based on these
- Effect of the DLC suppply temperature on the heat capture rate
- Effect of the rack temperature on the compute node power consumption and heat capture rate
- Interactions between rack temperature, DLC supply temperature on the heat capture rate and total power consumption
- Finding the best operating point, allowing as much free cooling as possible without increasing the compute node power consumption

Test conditions for cooling optimization of DLC systems

- Tests performed at 89kW (at initial cooling settings 27°C/27°C) per Rack (2 exclusively reserved racks with Firestarter and fixed clock speeds), measured at PDU
- Response surface split plot experimental design with additional star points for rack and secondary fluid circle supply temperatures (rack temperature is difficult to change parameter)
- Temperature range: DLC supply 22-32°C (star at 18°C, 34°C and 36°C) rack 24-30°C (star at 22°C, 32°C skipped as 30°C already too high for sensible operation)
- Heat capture rate of DLC based on CDU cooling duty estimates
- Heat capture rate of in-row cooler based on primary flowrate and ΔT through measured at balancing valve
- 10 to 20 min. waiting time before measurments to allow stabilization of values

Observations during operation and cooling experiments

DLC systems

- Reduction of DLC (secondary fluid circle) temperature increases heat capture rate (up to 90% at 18°C)
- CDU controller oscillations at $\Delta T > 8°C$ between primary and secondary circuit
- $\Delta T <= 4°C$ between primary and secondary supply temperature results in 100% CDU valve opening most of the time
- DLC return temperatures <40°C result in drop of CDU heat exchanger efficiency $\rightarrow$ lower primary $\Delta T \rightarrow$ reduced dry air cooler efficiency
- No cooling gain by increasing the in-row fan speed >90%
- Less than 65% in-row fan speed increases compute node temps
- Full free cooling (in-row) needs ΔT=9-12°C between rack and primary supply
- Compute chassis fan control has an exponential curve, rack temp >28°C large fan speed increase

Accuracy HCR estimate

CDU Primary Circuit:

```
 1st Qu.  Median    Mean  3rd Qu.
-2.5500  -0.2500 -0.6923  0.8500
```

Deviation from manual calculation based on in-row flowrate and ΔT small, but load dependent variation at fixed temperatures $\rightarrow$ makes no sense

CDU Secondary Circuit:

```
 1st Qu.  Median    Mean  3rd Qu.
  4.100    6.900   5.754   8.600
```

Larger deviation from manual calculation based on in-row flowrate and ΔT, but very stable estimate $\rightarrow$ useable with offset correction

Model HCR

```
lm(formula = HCR.Secondary ~ Real.S2n * Real.Airn + I(Real.S2n^2))

Residuals:
    Min      1Q  Median      3Q     Max
-8.6296 -0.6971  0.2849  1.0971  6.4345

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         88.5330     0.7507 117.932  < 2e-16 ***
Real.S2n            -4.5933     0.5997  -7.659 6.76e-08 ***
Real.Airn            1.2613     0.5439   2.319   0.0292 *
I(Real.S2n^2)       -1.1848     0.5522  -2.146   0.0422 *
Real.S2n:Real.Airn  -1.2686     0.5557  -2.283   0.0316 *

Residual standard error: 2.799 on 24 degrees of freedom
Multiple R-squared:  0.7176,Adjusted R-squared:  0.6705
```

Model rack power consumption

```
lm(formula = Power.20 ~ Real.S2n + Real.Airn)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3838 -0.1421  0.0273  0.1083  0.3643

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) 89.47299    0.03733 2396.819  < 2e-16 ***
Real.S2n     0.13927    0.03708    3.756 0.000881 ***
Real.Airn    0.19221    0.03703    5.191 2.03e-05 ***

Residual standard error: 0.1921 on 26 degrees of freedom
Multiple R-squared:  0.6838,Adjusted R-squared:  0.6595
```

Model fanspeed

```
lm(formula = Fan.20 ~ Real.Airn + I(Real.Airn^2)])

Residuals:
    Min      1Q  Median      3Q     Max
-723.70  -80.91   28.25  116.42  430.69

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     7744.42      56.40 137.301  < 2e-16 ***
Real.Airn        540.75      62.91   8.596 4.48e-09 ***
I(Real.Airn^2)    98.08      40.88   2.399   0.0239 *

Residual standard error: 241.4 on 26 degrees of freedom
Multiple R-squared:  0.7941,Adjusted R-squared:  0.7783
```

Logical consequences

Primary circuit and air cooled systems

- Dry air coolers have now 14°C minimum allowing power saving for more than 30% of the year.
- Each CRAC unit has individual start temperatures for its compressor circuits and zone control, resulting in a 7 step power control
- Cold aisle target temperature increased to 23.5°C (highest possible temperature to ensure cold aisle <=26°C without unnecessary compressor cycling)
- Dynamic CRAC fan speed control from 78% up to 87%

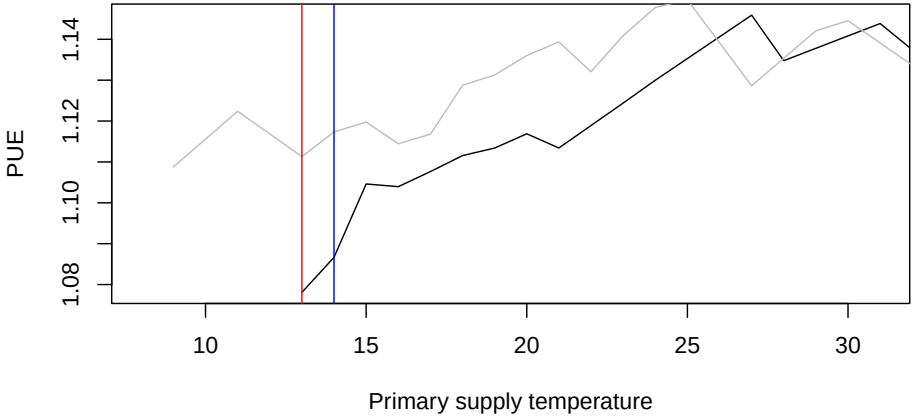
Logical consequences

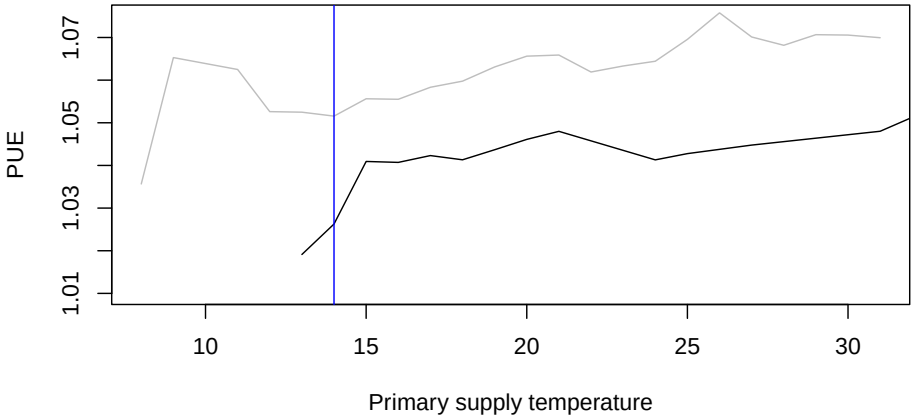
DLC systems

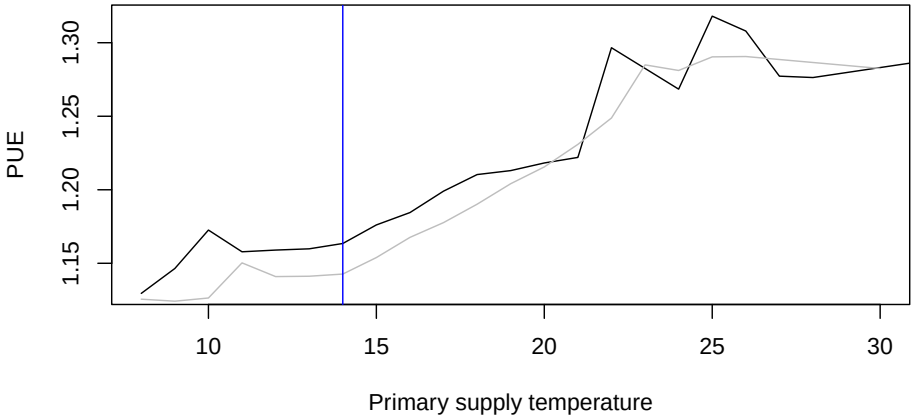
- Rack temperature target for DLC systems increased to 27.5°C
- Fan speed limits for in-row cooler fans now 65% and 90%
- Fine tuning of hydraulic balancing under partial load to achieve same free cooling capacity for all in-row coolers under all load conditions
  $\rightarrow$ full free cooling at up to 20°C primary supply (5°C higher than before)
- Dynamic control of DLC supply temperature following the primary fluid circle temperature with $\Delta T=5°C$, minimal temperature of 22°C and maximum of 34°C (resulting in 79-85% heat capture rate)
- Heavy tuning of CDU PID parameters to allow stable control with primary to secondary $\Delta T>8°C$ (standard parameter range was not sufficient so cooperation with Luc from CoolIT to provide patched firmware)

Outlook

Primary temperature	PUE year 1	PUE year 2 (first months)
13	1.111304	1.078584
14	1.117340	1.102545
15	1.119757	1.104047
16	1.114439	1.100661
17	1.116774	1.105555
18	1.128738	1.109466
19	1.131273	1.114417
20	1.136026	1.117766
21	1.139366	1.117321


PUE
1.08
1.10
1.12
1.14
10
15
20
25
30
Primary supply temperature


PUE
1.01
1.03
1.05
1.07
10
15
20
25
30
Primary supply temperature


PUE
1.15
1.20
1.25
1.30
10
15
20
25
30
Primary supply temperature