POLITECNICO
MILANO 1863

textarossa

The TEXTAROSSA Approach to Thermal Control of Future HPC Systems

Ver 1.0

William Fornaciari, - Politecnico di Milano - DEIB

https://textarossa.eu/

Outline of the talk

- The TEXTAROSSA project
- The TEXTAROSSA platform
    - Integrated Development Vehicles (IDV)
- Multi-level thermal management
- Assembling the technologies
    - Thermal test chip
    - 2-phase cooling
    - Event-based control
- Conclusions

The TEXTAROSSA Consortium

- Project coordinator: Massimo CELINO, ENEA
- Project Tech. Manager: William FORNACIARI, POLIMI
- Partners from 5 countries: ENEA, Fraunhofer, INRIA, ATOS, E4, BSC, PSNC, INFN, CNR, IN QUATTRO, CINI (Politecnico di Milano, Università di Torino, Università di Pisa), LTP: Universitat Politecnica de Catalunya (UPC), Université de Bordeaux
- EuroHPC Joint Undertaking, H2020 G.A. 956831
- Duration: 36 months (April 2021 – March 2024)
- Budget 6 M€
- Web site TEXTAROSSA: https://textarossa.eu/
- Website CINI lab on HPC: Key Technologies and Tools
    - https://www.consorzio-cini.it/index.php/it/laboratori-nazionali/hpc-key-technologies-and-tools

Main technical goals of TEXTAROSSA

1. **Energy efficiency and thermal control**
   - innovative two-phase cooling technology at node and rack level, fully integrated in an optimized multi-level runtime resource management
2. **Sustained application performance**
   - efficient exploitation of highly concurrent accelerators (GPUs and FPGAs) by focusing on data/stream locality, efficient algorithms and programming models, tuned libraries and innovative IPs
3. **Seamless integration of reconfigurable accelerators**
   - by extending field-proven tools for the design and implementation such as Vitis and OmpSs@FPGA to support new IPs and methodologies such as mixed-precision computing and power monitoring and control
4. **Development of new IPs**
   - for mixed-precision AI computing, data compression, security, power monitoring and control, and scheduling
5. **Integrated Development Platforms**
   - by developing two architecturally different, heterogeneous Integrated Development Vehicles (IDVs), one as a dedicated testbed for two-phase cooling technology, and one supporting the wider range of project technical goals

Where is the “power” and “thermal” to be managed in TEXTAROSSA?

- The computing architecture are heterogeneous, several contributions to the power
  - CPUs, GPUs and FPGA-based hardware accelerators
  - Organization in blades and racks
- Why controlling power and thermal aspects?
  - Reliability, availability, green computing, cost, …
  - …, it is mandatory to scale-up the computing power of exascale centers
- Aspects to be considered
  - Time scales of phenomena (tens of ms, seconds, tens of seconds or more, …)
  - A priori knowledge of the computing needs
  - Monitoring of power, which temperature is relevant, granularity of the analysis
  - Knobs, actuators (DVFS, task migration, stop&go, …)
  - Proactive vs Reactive run-time power management policies
  - Impact on the workload execution
  - Cost and complexity of the cooling
  - Coexistence of several local control loops

- **In TEXTAROSSA, concerning the thermal aspects, it is under development**

- Physical and mechanical design of the rack mounted system, that is not trivial

- Thermal models and power monitors for different types of elements
  - CPUs, GPUs
    - Some similarities
  - FPGA accelerators
    - Not standard, functionality and power patterns can change dynamically, need to generate ad-hoc power monitors

- Focus of the talk: the “cool” cooling of the future is 2-phase cooling!
  - Improvement of the cooling efficiency up to 70% compared to air cooling, up to 30% compared to liquid cooling
  - It will be tested on both ATOS and E4 infrastructures (IDVs)

- First steps in thermal modeling and development of multi-level control

Integrated Development Vehicles (IDVs)

- Two architecturally different, heterogeneous Integrated Development Vehicles (IDVs) will be developed
  - IDV-A by ATOS, X86/64 and GPUs
  - IDV-E by E4, featuring ARM and FPGA

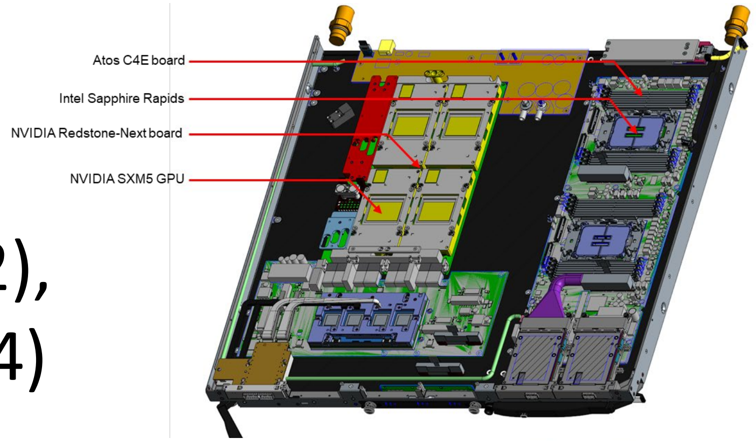
- These IDVs will be used as testbed and workhorse by TEXTAROSSA's developers
  - The IDVs will be a single-node platform, easy to configure and reconfigure, extensible in terms of components, devices, peripherals, flexible in terms of supported SW (OS, utilities, drivers, runtime libraries), instrumented with thermal sensors, electric probes, thermally induced mechanical stress sensors
  - The developers will use the IDVs to test their codes, algorithms, drivers without having the constraints of a large system and having the advantage to be able to test their developments on different components through a very quick reconfiguration process

Integrated development vehicles IDV-E (E4), IDV-A (Atos)

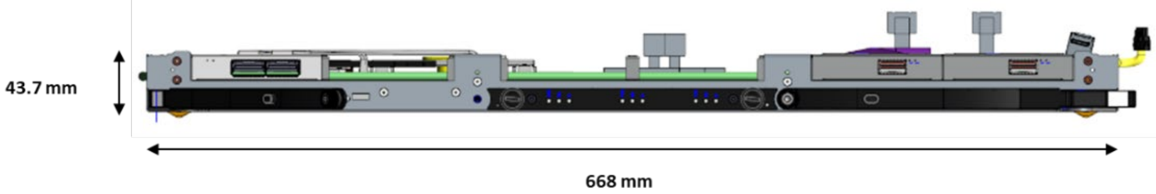
IDV-E – E4

- 2xU280 Xilinx FPGA (TDP 225W), 2xAmpere Altra Max processor (based on ARMv.8.2+) offers up to 128 cores operating at a maximum of 3.0 GHz, 7 nm FinFET with a TDP of 250 W
  Total thermal power of computing: **950W**

IDV-A – Atos



- CPU: 2-socket Intel Sapphire Rapids (TDP 350Wx2), connected to 4 Nvidia Hopper GPUs (TDP 700Wx4)
- Total thermal power of computing units: **3.5 KW**



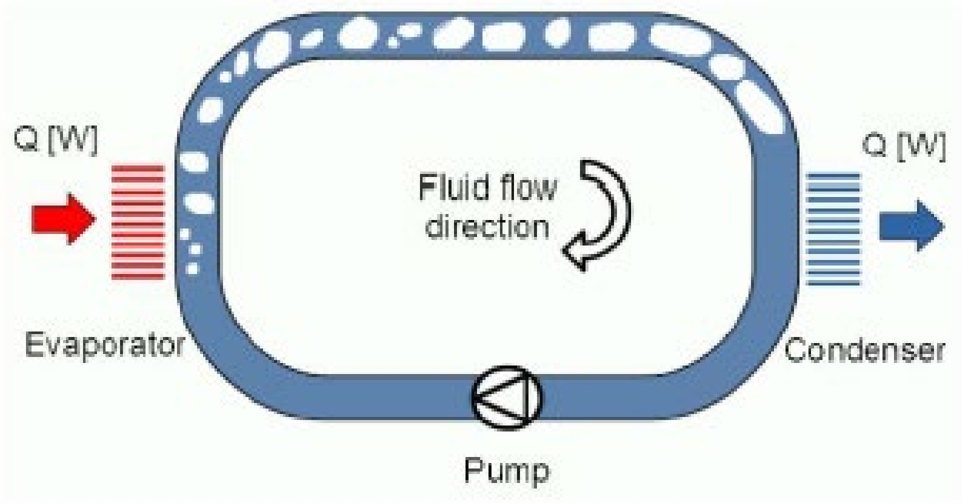
2-phase cooling by IN4 installed Oct 16-20, @Atos first week of November, 2023

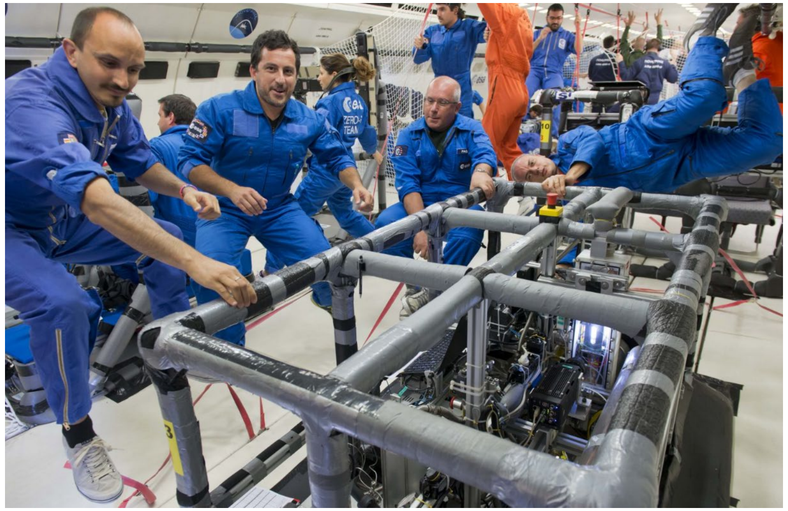
IDV-E by E4 (born a pair of weeks ago): Mt. Collins System

- New integrated heterogeneous architecture at node level
  - Integration of EPI experience (processor and accelerators IPs, and integrated heterogeneous HPC platforms exploiting both ARM64 and RISC-V cores) to boost the EuroHPC roadmap in terms of energy-efficiency, high-performance and secure HPC
  - U280 Xilinx FPGA (225W max), Ampere Altra Max processor (based on ARMv.8.2+) offers up to 128 cores operating at a maximum of 3.0 GHz, 7 nm FinFET with a TDP of 250 W
- Room to host new **two-phase cooling technology** at node and system levels, power monitoring and controller IP exploiting new models of the thermal behavior and of a multi-level control strategy
- Innovative tools for seamless integration of reconfigurable accelerators: such tools, targeting the AI/DNN computing paradigm, include compilers, memory hierarchy optimization and runtime systems, scaling over multiple interconnected reconfigurable devices, and SW header-only based on Fast Flow and memory hierarchy optimization in an EPI-like HPC architecture, compiler tools for mixed-precision, all in heterogeneous HPC platform and in future EPI tool chain
  - Automatic instrumentation of the accelerators with energy/power models to enhance a global (fine grain) power monitoring and control

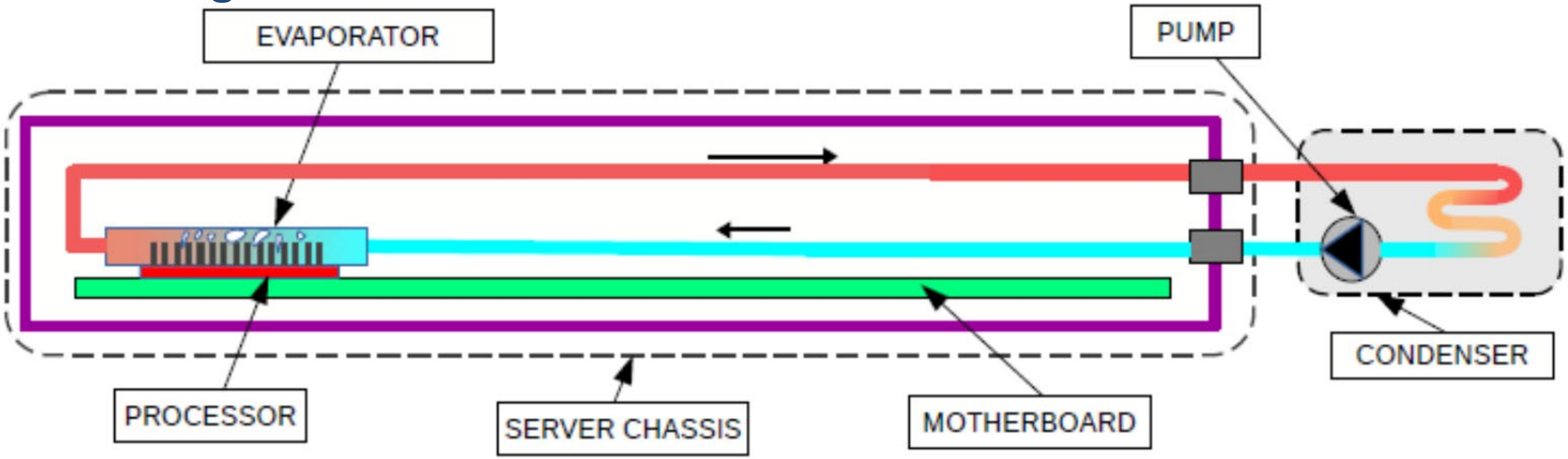
2-phase cooling principle

- Compared to classical cooling systems like heat pipes and liquid cooling (single phase forced convection), the new technology is able
    - to remove higher heat fluxes, at lower pumping energy, lower mass of the entire loop, and to maintain the target surface of the electronic component isothermal
- Cold plate, (aluminum or copper), that serves as a heat sink, a direct on chip **evaporative** heat exchanger
- The cold plate is a multimicro-channels evaporator
    - a metal plate with micro fins machined on it





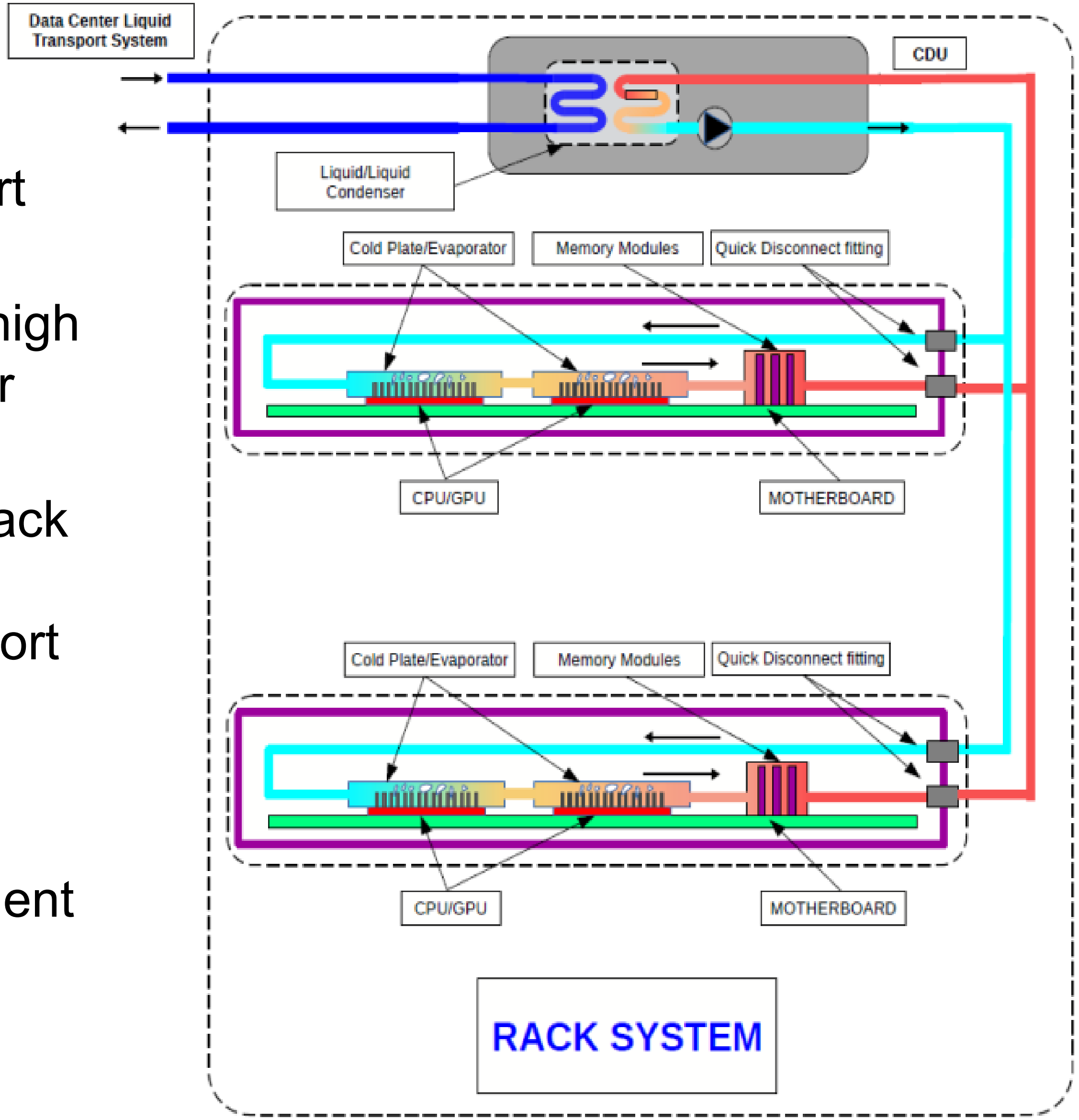
2-phase cooling for CPUs/GPUs



- The evaporator is placed in direct contact with the processor (CPU or GPU)
- The coolant flows through micro-channels in the evaporator to capture heat from the processor by evaporation, and then flows on to a condenser, in which heat is dissipated to the surrounding environment via water or air
- The coolant coming out of the condenser travels back through the pump, and the cycle repeats itself
- The loop is a hermetically sealed closed system, so the processors and all electronic components are not in direct contact with the fluid
- The use of dielectric liquids eliminates the risk of electric damages caused by an accidental leakage of the coolant
- Such dielectric fluids are non-flammable, non-toxic, perfectly compatible with the environment with extremely low GWP (Global Warming Potential) and Ozone Depletion Potential (ODP)

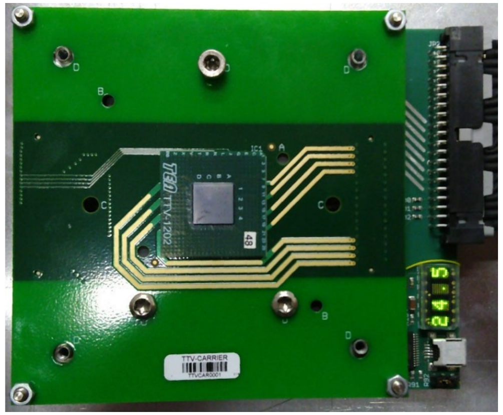
Moving up to system level

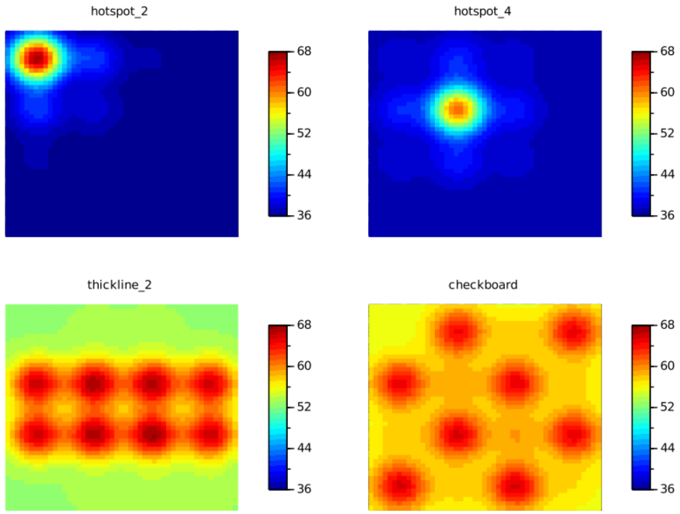
- The two-phase cooling system will be integrated at rack level for data center with liquid thermal transport infrastructure installed
- The two-phase cooling system will be designed for high density CPU and GPU configurations up to 5 kW per server
- All the server cooling systems will be integrated at rack level with a CDU (Coolant Distributor Unit) that will transfer heat to the data center liquid thermal transport infrastructure
- Each rack will be able to remove up to 90 kW using water in the liquid transport infrastructure as hot as 45°C, eliminating the need for expensive and inefficient chillers or cooling towers

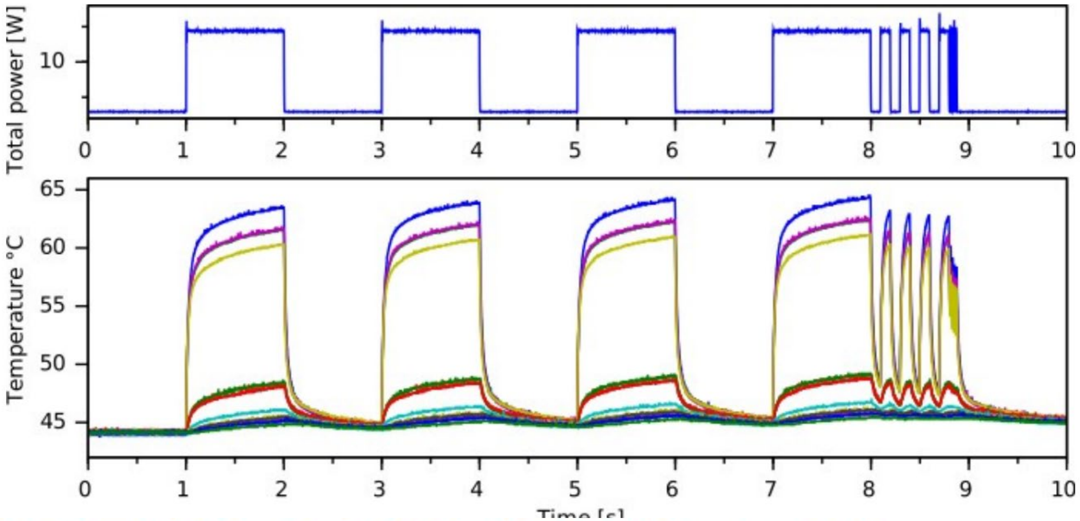


Thermal Test Chip (TTC): experimental campaign to create a thermal model supporting multi-level thermal control

- A TTC is a custom BGA IC (die area 10.23mm x 10.23mm) with
  - A 4x4 array of power sources, driver speed 300us
  - An array of temperature sensors with 0.1°C resolution
  - 12Wx16=192W max power, 150°C max T
- Provide a way to measure silicon temperatures and their spatial distribution
  - Can connect an arbitrary heat sink type ➔ Unlike IR thermography which forces the use of an oil bath
- Useful for designing thermal models of heat sinks

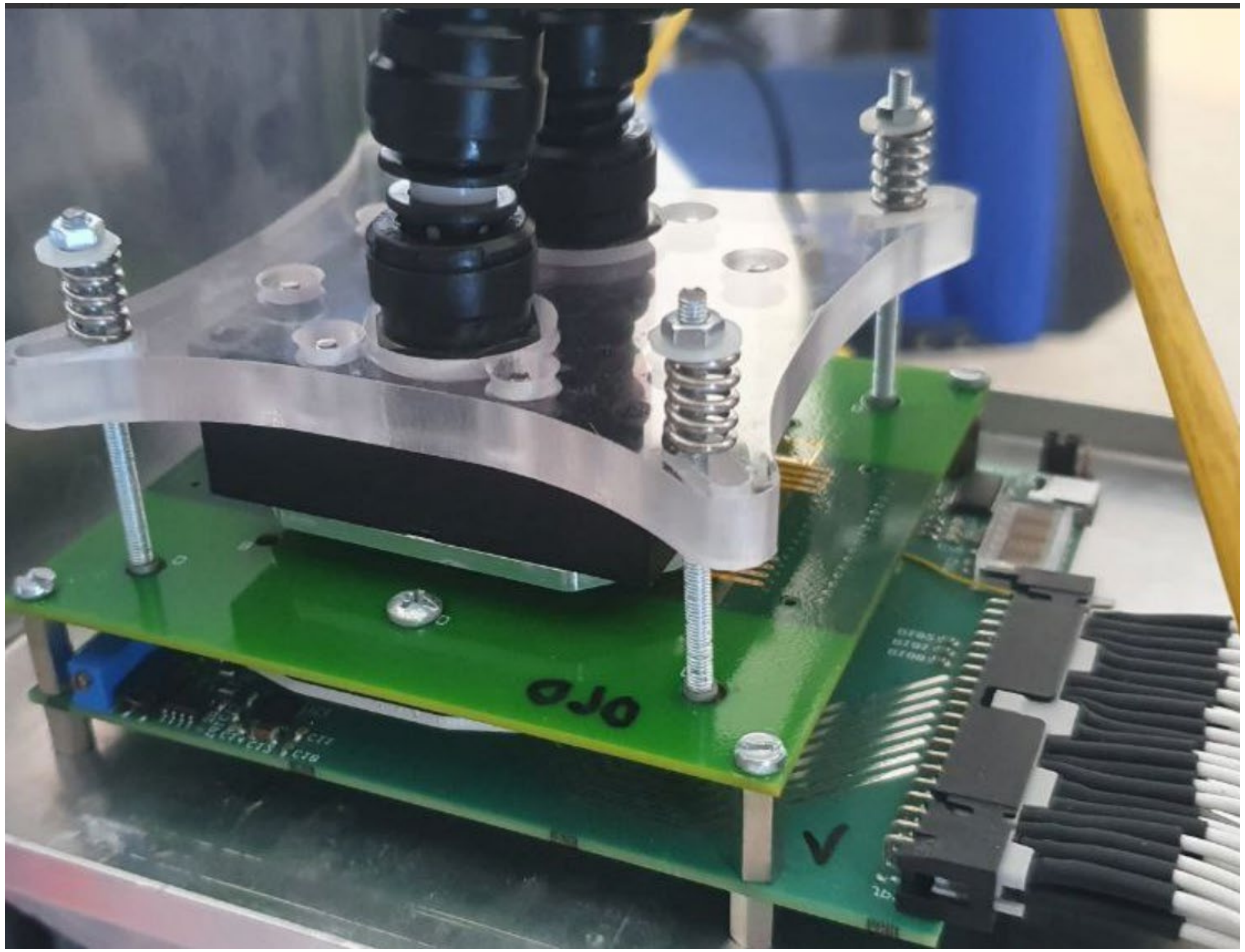


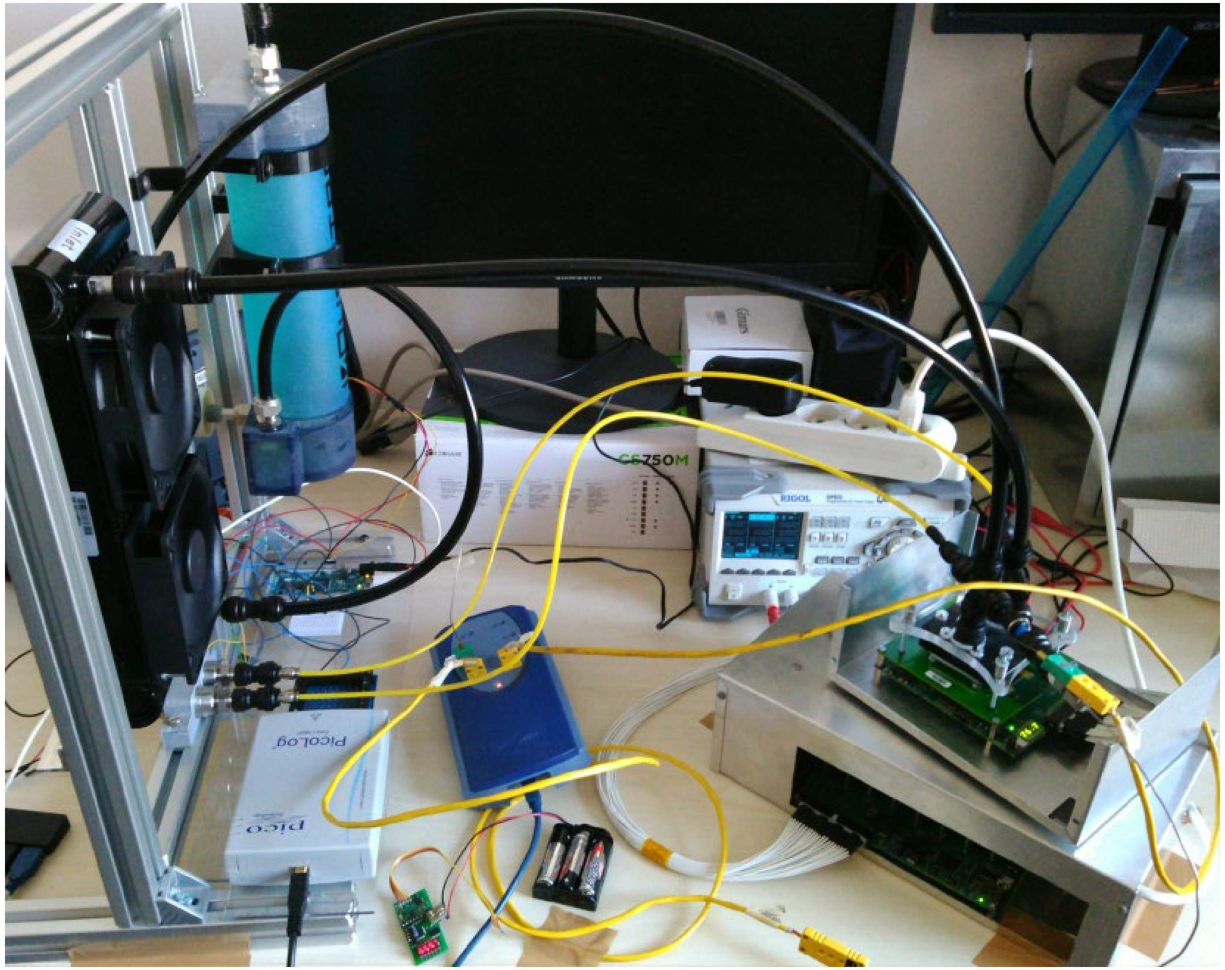




https://link.springer.com/chapter/10.1007%2F978-3-030-27562-4_13

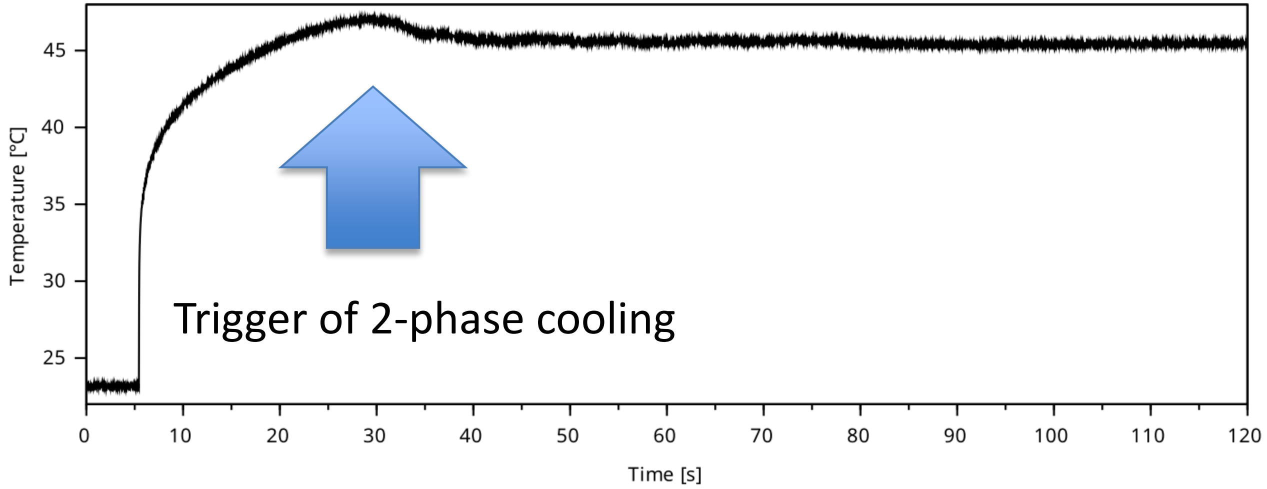
Experimental campaign (IN4-POLIMI) to create a thermal model





Preliminary validation and WiP

- **It works !** Intervention of 2-phase cooling is effective, no temperature overshooting
  - Still using old built-in control loop

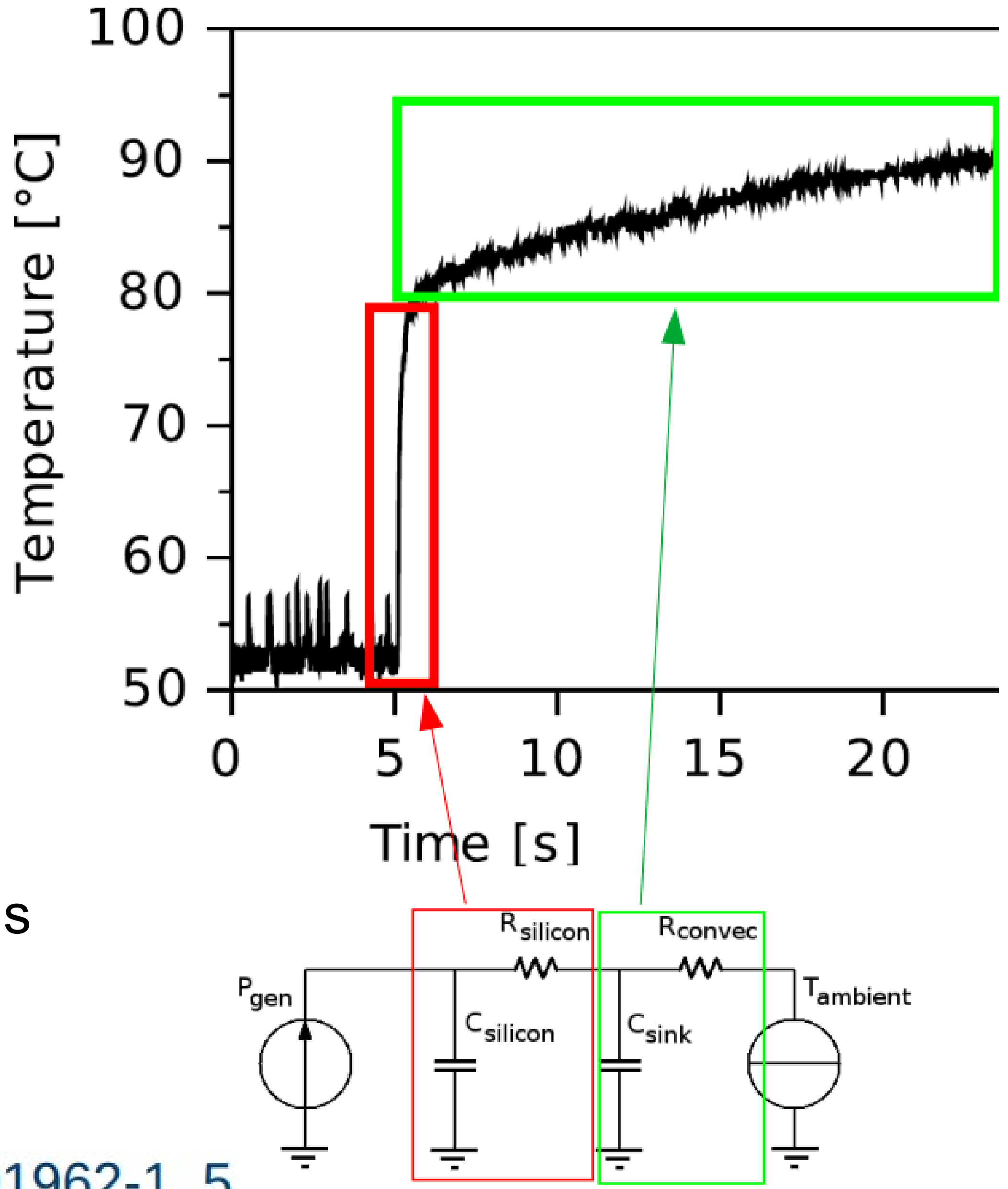


Trigger of 2-phase cooling

- 2-days of collection of measurements to develop an Object Oriented (Modelica) model for this “type” of dissipation
  - Thermal simulation will be possible
  - Spatial distribution of temperature over the chip could be considered
  - Event-based thermal control will be integrated and customized to work in synergy with 2-phase cooling and DVFS

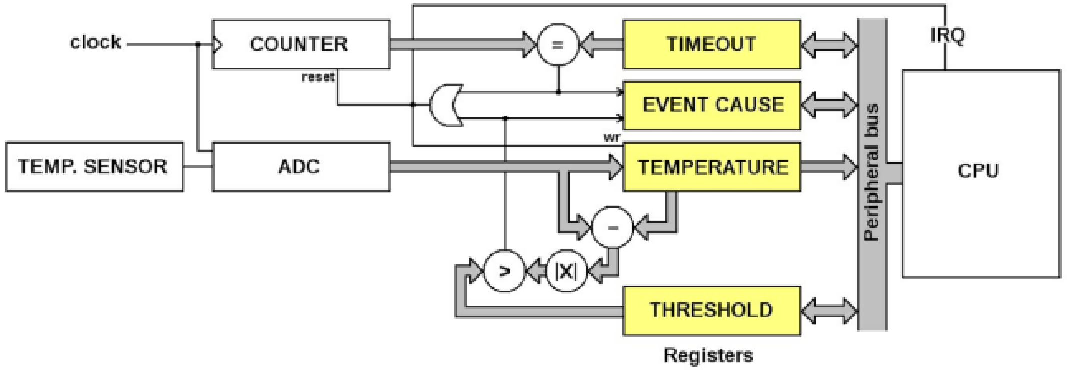
Modeling thermal phenomena

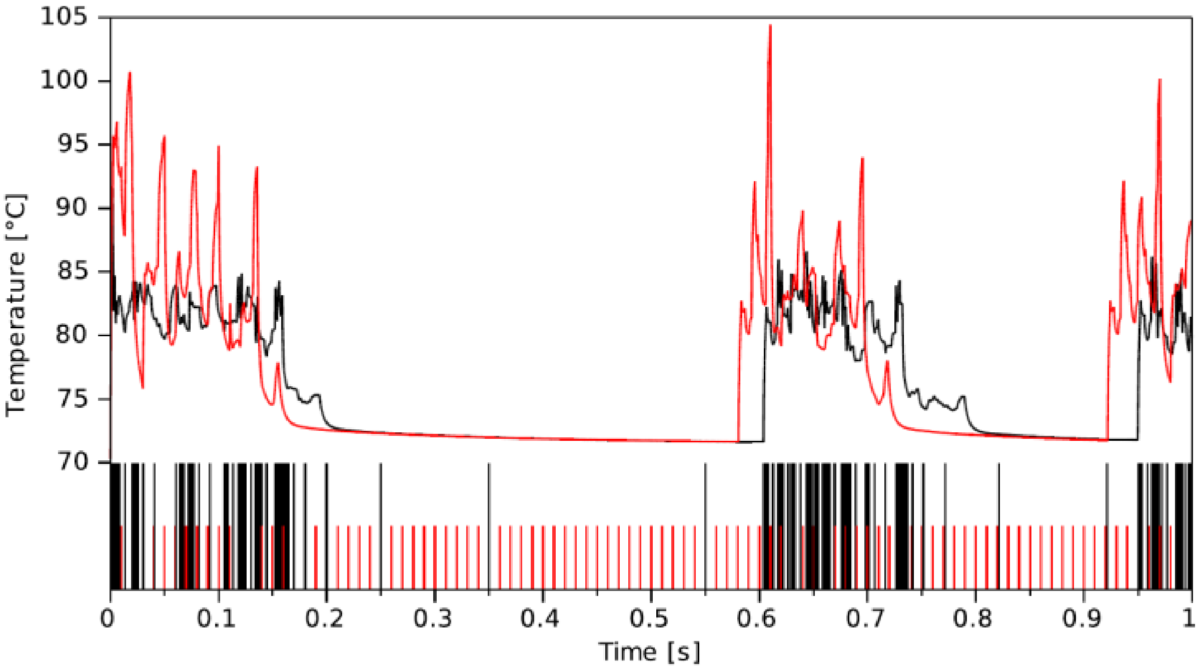
- Experimental setup
  - Intel Core-i5 6600K
  - Instrumented Linux kernel for sampling CPU temperature at 1KHz
  - cpuburn program run @ t=5s
- Thermal dynamics
  - Baseline temperature 54°C (noisy)
  - Initial temperature rise caused by fast thermal dynamics
  - 70°C reached in 50ms
  - 80°C reached in 600ms
  - Subsequent temperature increase happens due to the heat sink thermal dynamics
  - 90°C reached in >20s



https://link.springer.com/chapter/10.1007/978-3-319-91962-1_5

Event-based thermal control (US Patented)





- The policy is not executed at a fixed rate, but when events occur
  - Based on a hardware-software codesign
  - A small hardware state machine monitors sensors and generates events
- Events cause the execution of a software control policy based on control theory
- Faster response with less overhead

US Patent granted Nov. 2021 - US11163345B2 https://ieeexplore.ieee.org/document/7890422

Conclusions

- The TEXTAROSSA project aims to achieve a broad impact on the HPC field both in pre-exascale and exascale scenarios
- The TEXTAROSSA consortium will develop new IPs, algorithms, methods and software components for HPC-AI, HPC and HPDA applications, mostly Open Source and able to be adopted as standalone building blocks or to interoperate with other Exascale-ready components

- Innovative **two-phase cooling** system for node(s) with the objective to serve an entire rack
    - Improvement of the cooling efficiency up to 70% compared to air cooling, up to 30% compared to liquid cooling
    - Testing on real platform prototypes (IDV) industry-grade
- Design and validation of thermal models taking advantage of equation-based object-oriented modeling languages to easily account for two phase liquid cooling
- Model validation uses a thermal test chip to capture accurate thermal maps of chips connected to the proposed heat dissipation solution
- Multilevel control allows to partition the system level control problem into multiple interacting control loops, each optimized for the specific thermal dynamics to control

Thanks for your attention – POLIMI research group

Politecnico di Milano – DEIB main staff for TEXTAROSSA

Prof. William FORNACIARI, Giovanni AGOSTA

william.fornaciari@polimi.it, agosta@acm.org

Prof. Federico TERRANEO, Prof. Davide ZONI, Dr Federico REGHENZANI, Dr Giuseppe MASSARI, Dr Andrea Galimberti

name.surname@polimi.it

Our Lab on Embedded and HPC computing

HEAPLab: http://heaplab.deib.polimi.it/

Active Projects on HPC / Computing Continuum

- **Textarossa (Euro-HPC, 2021-2024): https://textarossa.eu/**
- ITN «Apropos» – ITN on Approximate computing: https://projects.tuni.fi/apropos/

Other EURO-HPC projects started in early 2022 (technology and application Pilots)

- **The European Pilot - EuroHPC**
- **Eupex – EuroHPC**
- **ISOLDE - KDT**