SIPEARL

Evaluation and performance projections for ARM chips
AHUG workshop / ISC 2023

Etienne Renault

SiPearl, the company born from a European will

September 2018

EuroHPC Joint Undertaking

Launch of the EuroHPC Joint Undertaking
backed by a €8bn budget
to deploy in Europe a world class exascale
supercomputing infrastructure

December 2018

European Processor Initiative

Launch of a call for proposals in 2017 for developing
a new generation of high-end European microprocessors

- Budget: €150m
- Target: high-performance and energy-efficiency

Coordinated by Bull (Atos Group), the European Processor Initiative (EPI) consortium won this call for proposals. It has currently 28 members:

- Scientists: research institutes, universities and supercomputing centres
- Industry: European leaders, IT, electronics and automotive specialists

June 2019

SiPearl is the private company created within the EPI
to launch a strategic industry for Europe.

SiPearl in a nutshell

Building the world first energy-efficient HPC-dedicated microprocessor designed to work with any third-party accelerator (GPU, artificial intelligence, quantum).

Incorporated in June 2019

Key personnel from
Atos, MediaTek, ST, Marvell, NXP, Infineon, intel, NOKIA

Funded by the European Union

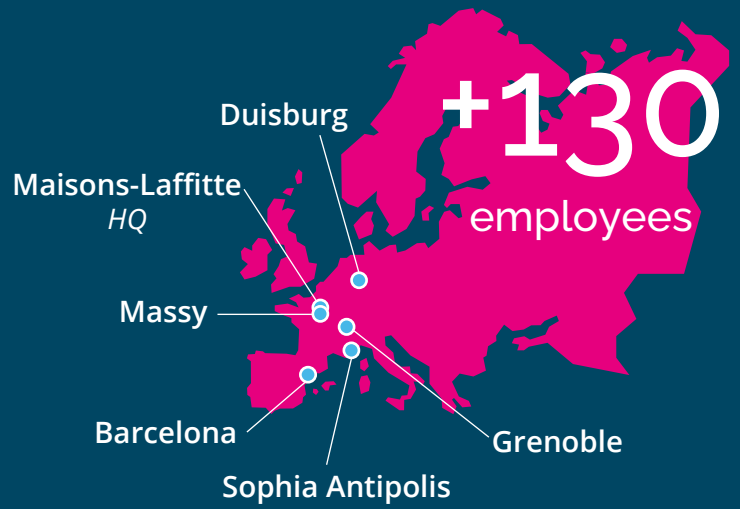
Financing
Initial closing of the Series A: €90m
arm, European Innovation Council Fund, Banque européenne d'investissement, EVIDEN, France 2030

ARM architecture
Energy-efficiency, quick time to market, proven ecosystem

Fabless
Manufactured by TSMC, mandatory at this technological level

Identified customers
Server manufacturers based on user specifications (governments, supercomputing centres, academics, industries, etc.)

6 locations in Europe



Motivation

Have insights from

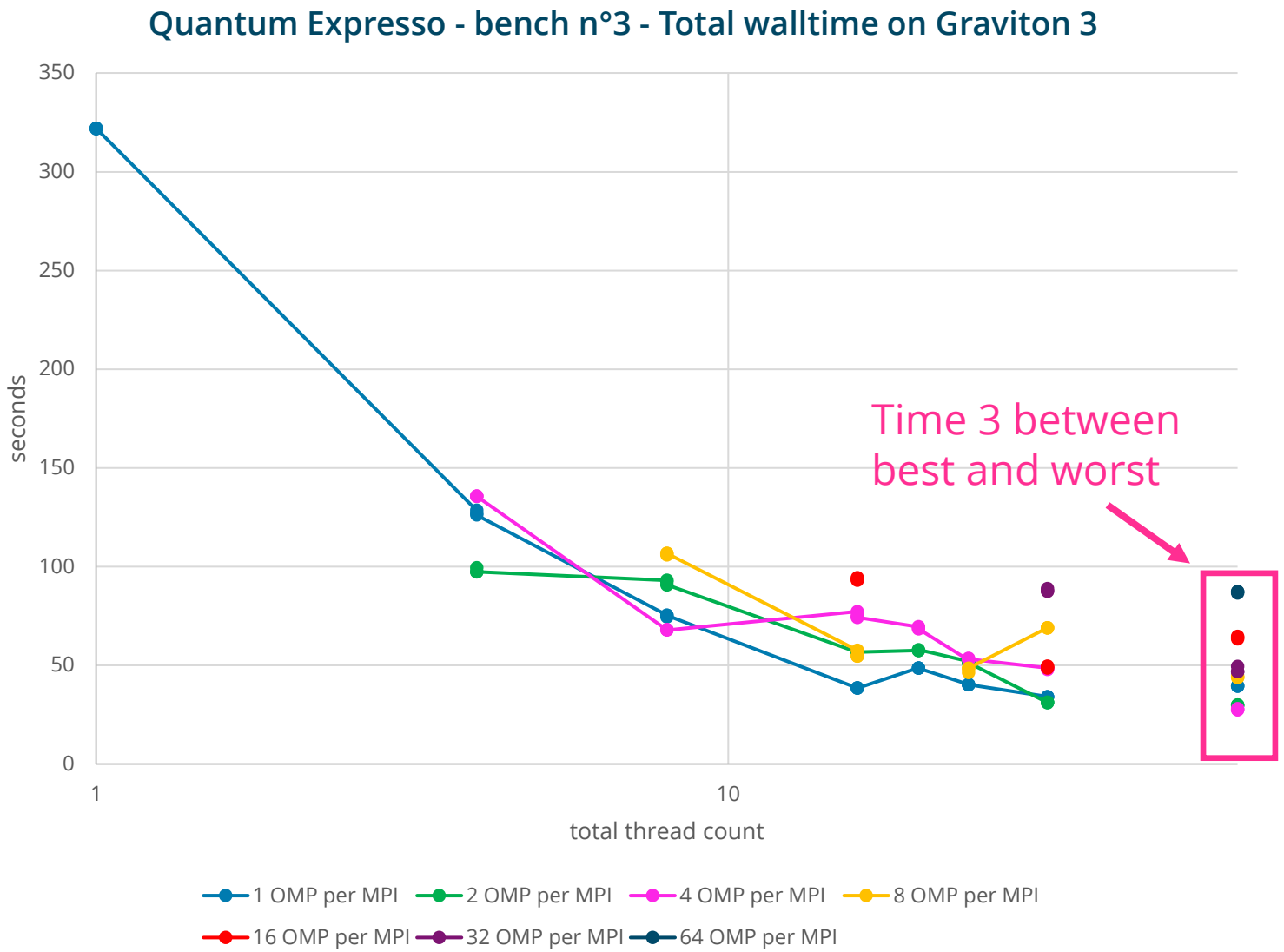
- What can be achieved from a given benchmark ?
- …On a given chip

Avoid brute force by anticipating the best combination to use

- Machines
- MPI processes
- OMP threads
- Numa nodes
- …

Build new chips

- SiPearl et al. job



Projection considered 1/2

All presented metrics are output agnostic

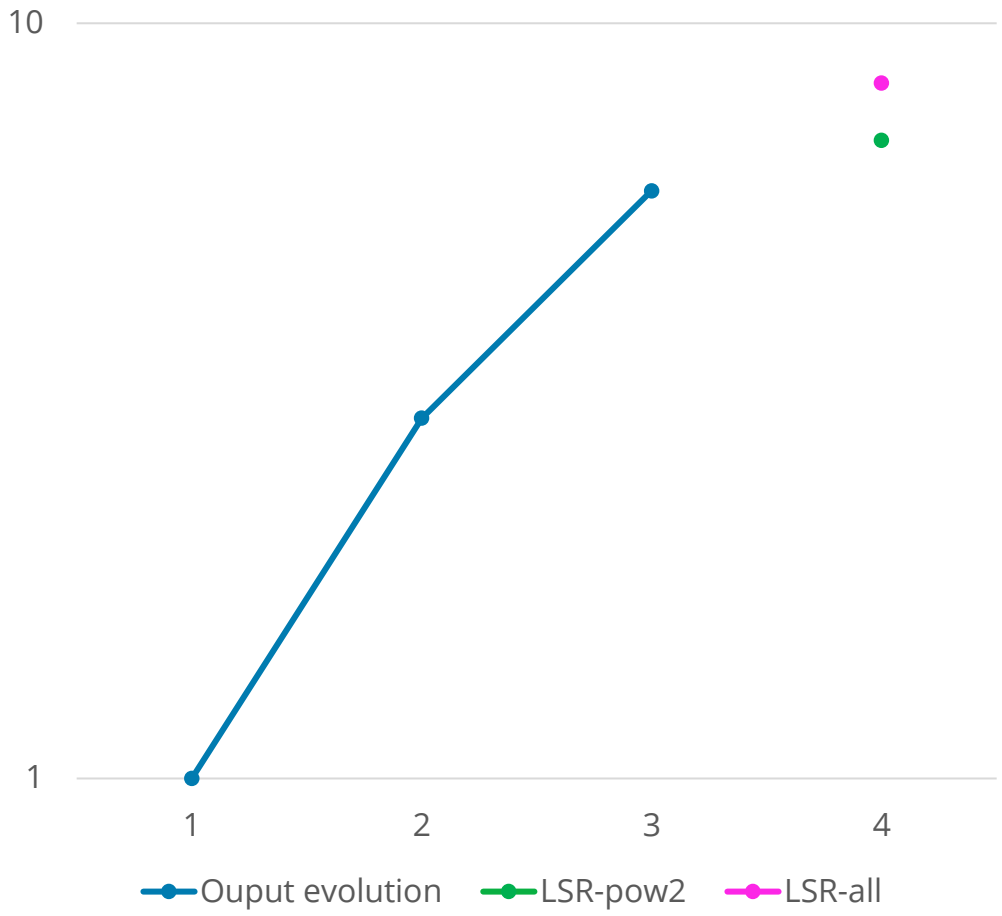
Least Square Root (LSR)

Compute next value (y) using $y = a + b.x$ using

$$a = \bar{y} - b.\bar{x} \qquad b = \frac{\sum(x - \bar{x}).(y - \bar{y})}{\sum(x - \bar{x})^2}$$

Two variations

- **pow2**
  - Baseline : the output for the 2 last power of two computations
- **all**
  - Baseline : the output for all the previous computations



LSR-all(4) = 8.3

LRS-pow2(4) = 7

Projection considered 2/2

Thread Impact (TI)

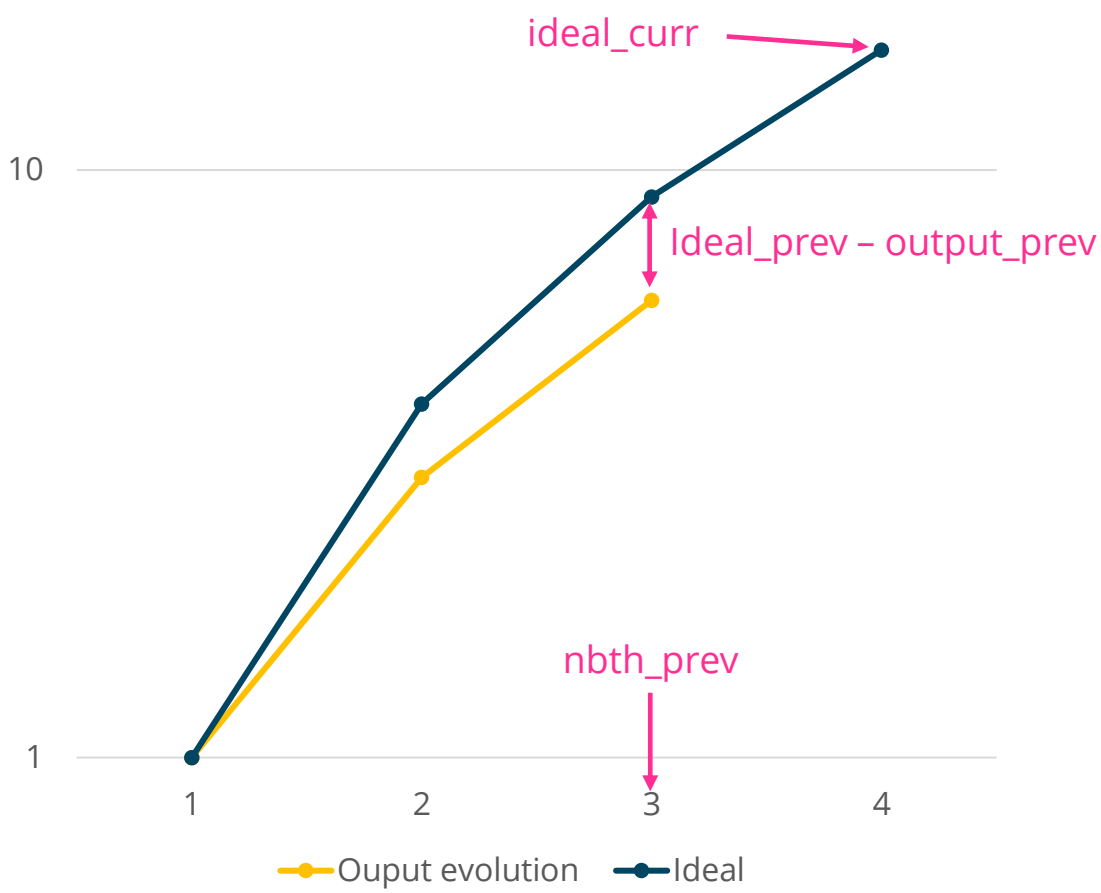
- Compute the « cost » of a thread at the previous computation
- Project it on the current computation

$$ideal_{curr} - \left(\frac{ideal_{prev} - output_{prev}}{nbth_{prev}}\right) * nbth_{curr}$$

Performance Drop (PD)

- Compute the output with the assumption that performance follows the one of another machine (here Ampere)

$$ideal_{curr} - ideal_{curr} * (1 - ratio_{ampere})$$



First case study : SPEC CPU

Benchmark description :

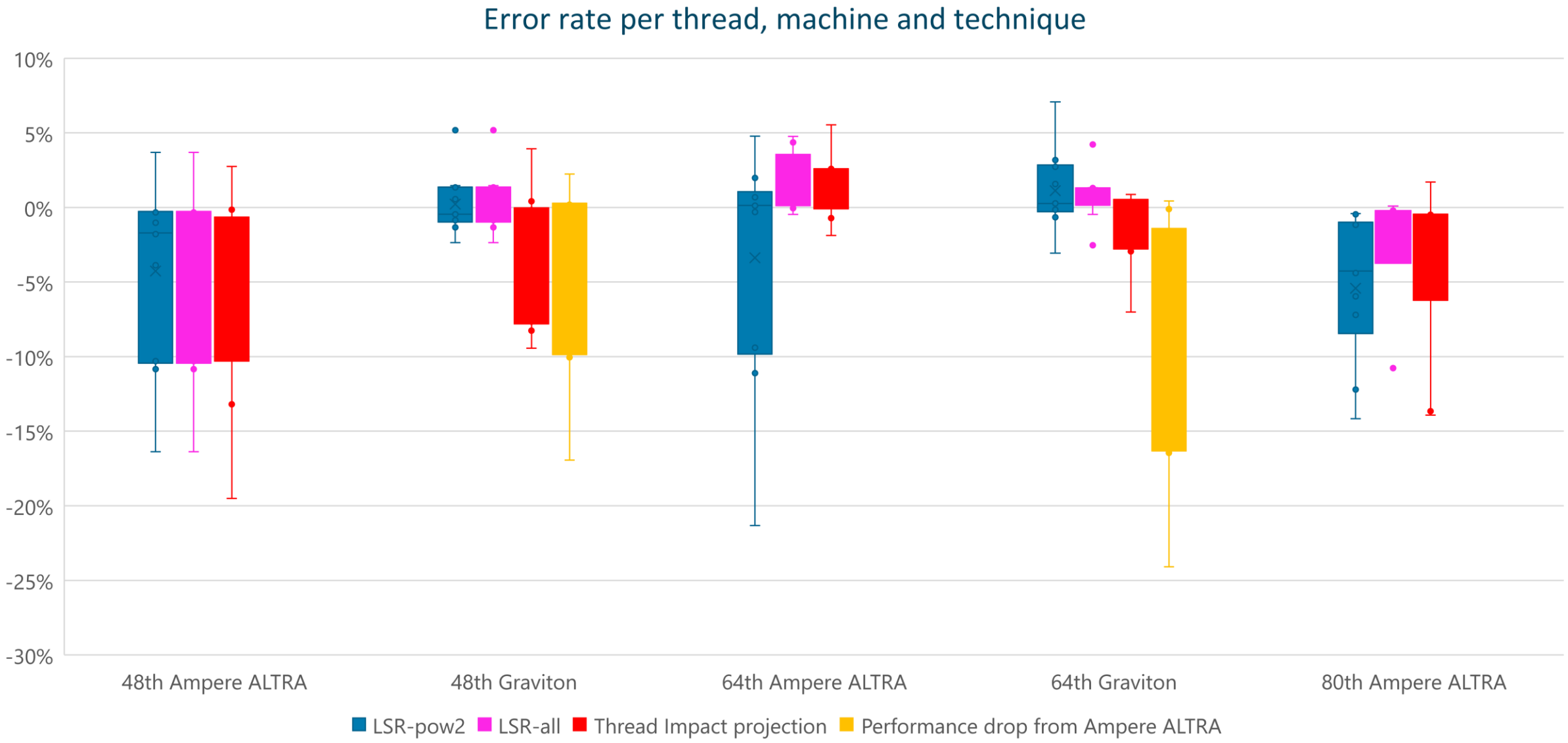
- Designed to provide performance measurements that can be used to compare compute-intensive workloads on different computer systems
- Run n copies of the underlying benchmark
- Expect no performance drop

Provide INT and FLOAT versions

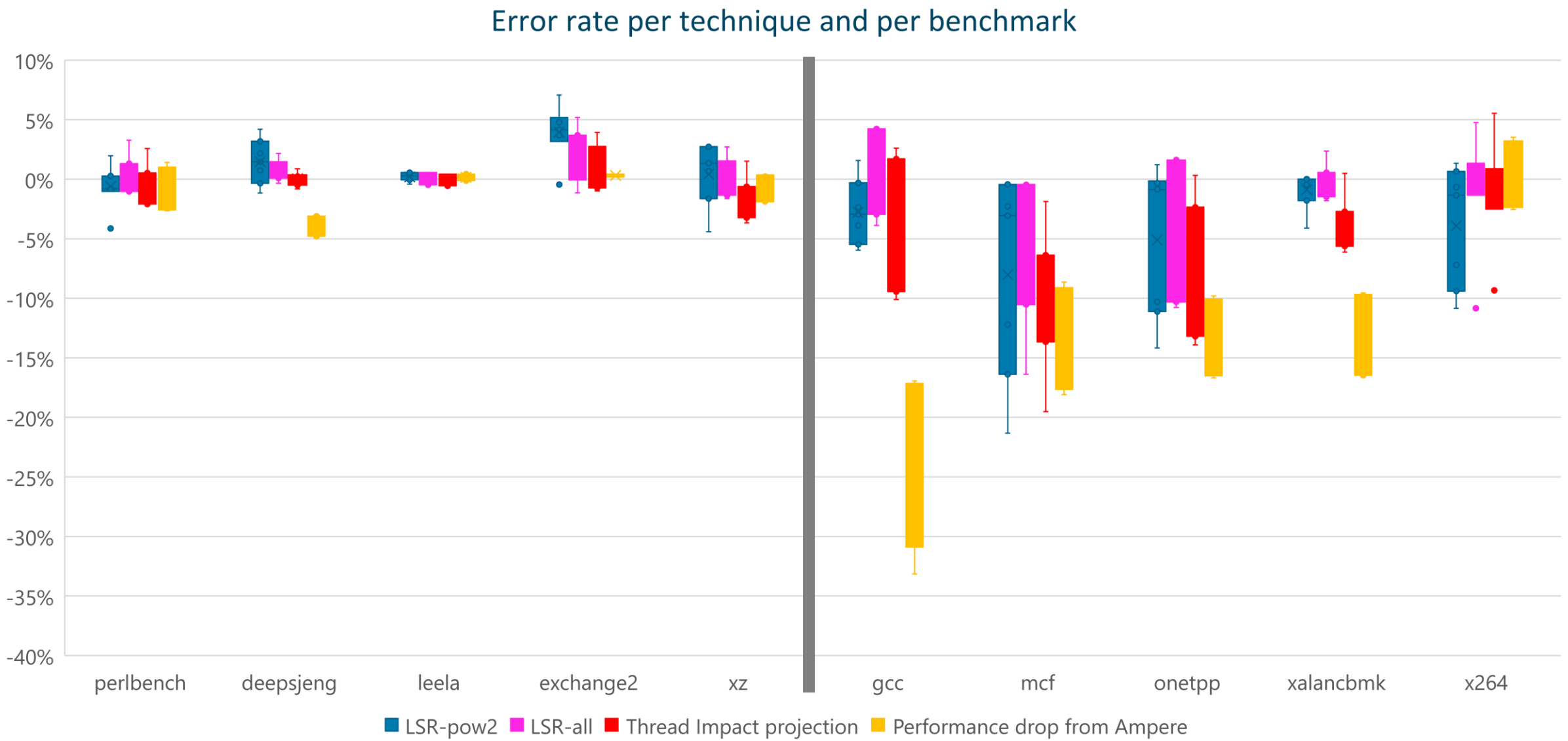
- Here only present the INT version without loss of generality
- Conclusions on the FLOAT versions are the same

Ideal output: stable whatever the number of considered threads

SPEC/CPU – Error Rate Summary 1/2



SPEC/CPU – Error Rate Summary 2/2



Partial conclusion

Walltime projections tends to be …

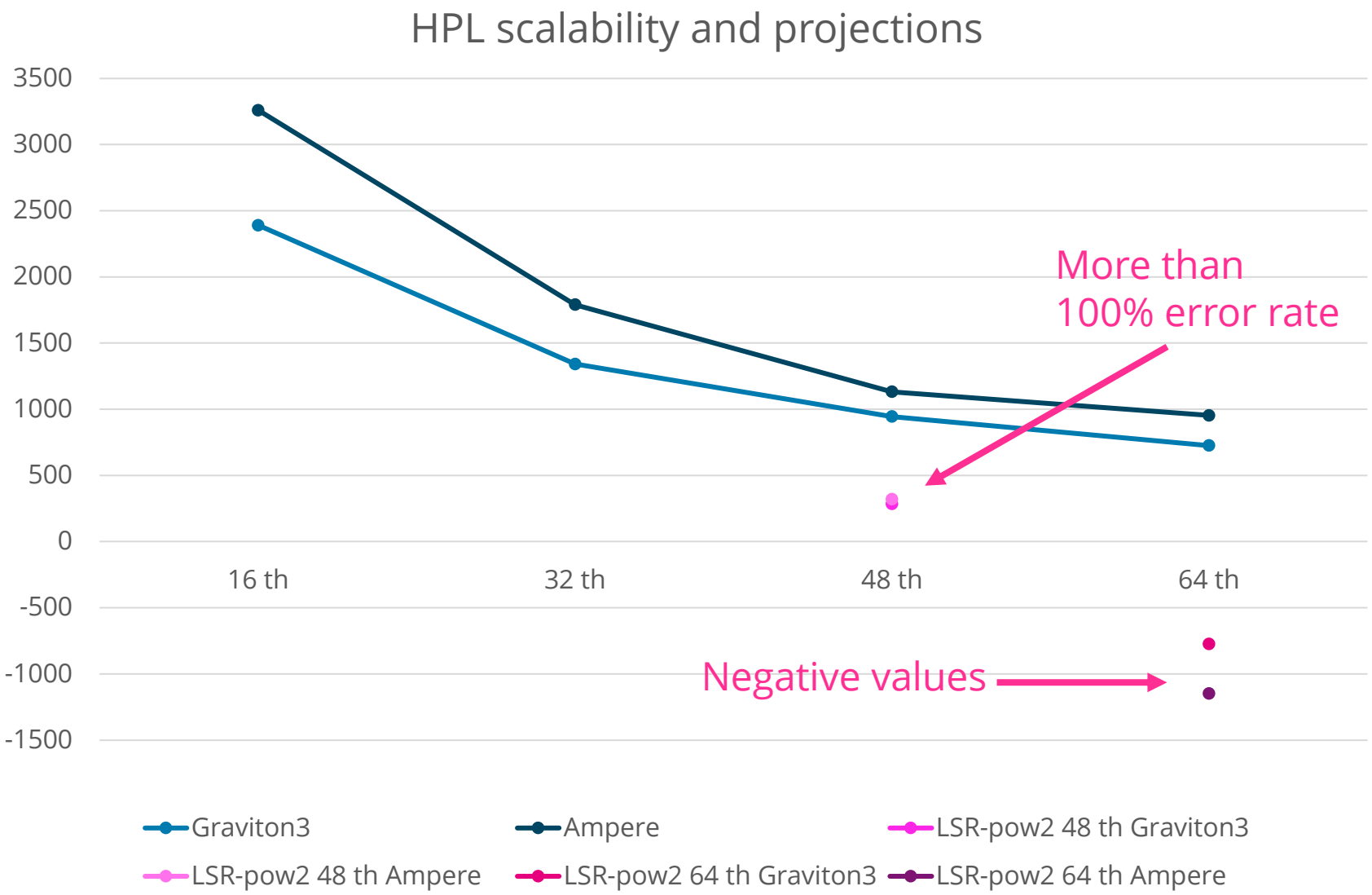
- … always better than reality

Projection type	Mean Error Rate
LSR-pow2	-2%
LSR-all	-1%
Thread Impact	-3%
Performance drop	-7%

- … in average better as the number of threads increase
- … sensitive in micro variation : projections must be done on averages
- … sensitive to bandwidth ⟹ need to anticipate correctly these projections

Non linear growth / decrease must be handled

The case of HPL – LSR-pow2 Projections



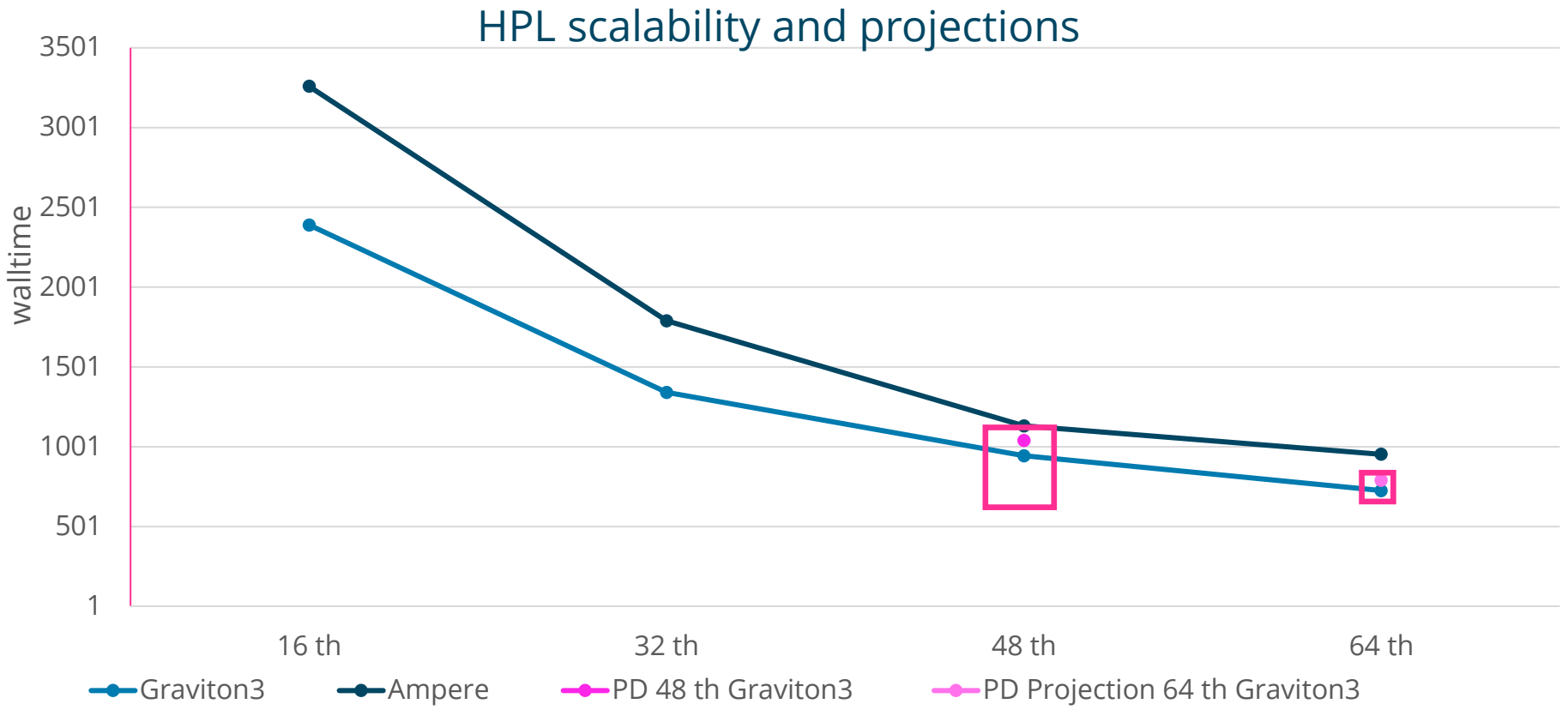
Similar results for LSR-all projections walltime is to sensitive

→

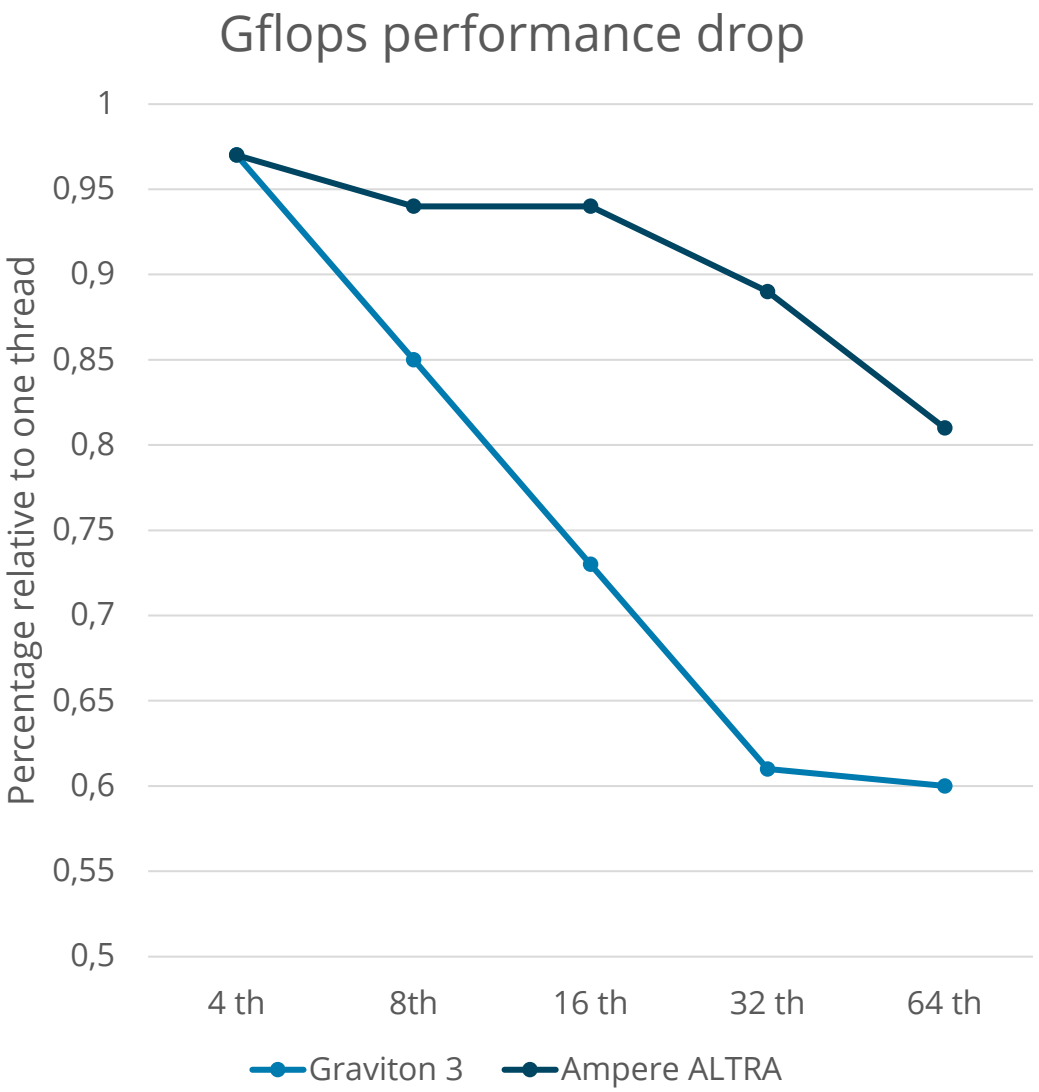
Need to pass inflexion point

The case of HPL – TI / PD

- For SPEC the **ideal** was Easy to compute, but her we don't have such metric.
- Will consider **ideal** as perfect scalability for walltime.



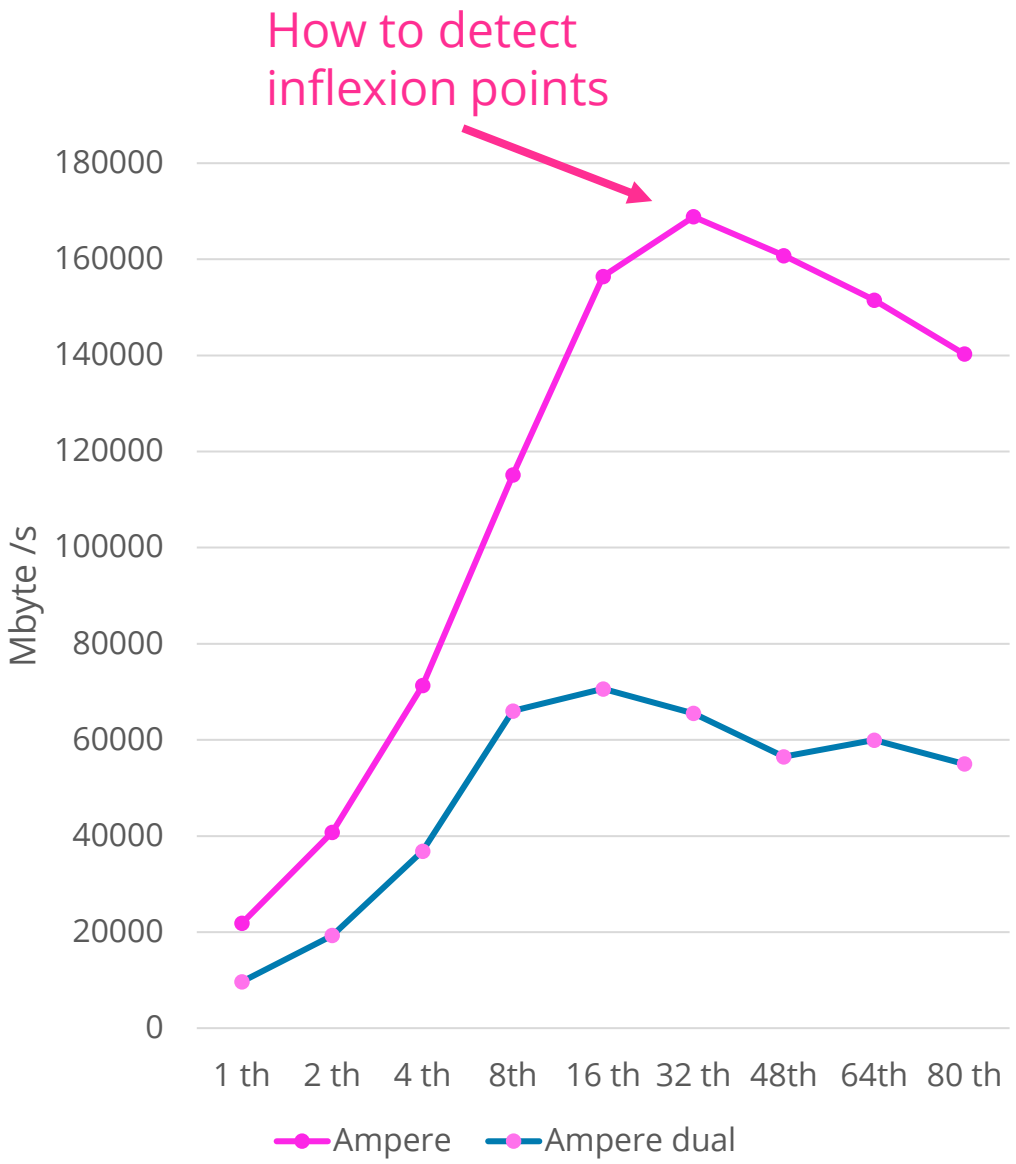
HLP – GFlops drop metric



More abstract metrics (like Gflops or ratio to peak) also work with projections

Projection type	Error Rate 48th On Graviton3	Error Rate 64th on Graviton3
LSR-pow2	2%	2%
LSR-all	5%	3%
Thread Impact	5%	3%
Performance drop	14%	-20%

STREAM - anticipate bad projections



	Metric	Error Rate 48th	Error Rate 64th	Error Rate 80th
Ampere single numa	Gflops	11%	22%	2%
	Walltime	-14%	-33%	-4%
Ampere dual numa	Gflops	7%	-6%	-4%
	Walltime	7%	-8%	4%

Since Gflops and time are strongly correlated, it is easy to detect incorrect projections

Case 48 th compared to 32 thread ;

- 16% faster **BUT** with 14% performance drop
- **INCONSITENCY**

→ Proj Gflops 0.08 rectification error rate = -6%
→ Proj walltime 0.14 rectification error rate = -3%

Conclusion

On the importance of combining multiple metrics

- To leverage erros and bypass local extremum
- To obtain one valid projection

About Projections

- LSR seem relatively stable, regarless the basline
  - But work better on other metrics than walltime
- TI/PD suitable for benchmark with inflexion points
- PD not adapted due to NOC differences
- Projections are working even better for Graviton3e

Future Work

- Tests projections with AFX64
- Run projections on HPC cluster, not single chips

About... SiPearl

SiPearl is building the world first energy-efficient HPC-dedicated microprocessor designed to work with any third-party accelerator (GPU, artificial intelligence, quantum). This new generation of microprocessors will first target EuroHPC Joint Undertaking ecosystem, which is deploying world-class supercomputing infrastructures in Europe for solving major challenges in medical research, artificial intelligence, security, energy management and climate while reducing its environmental footprint.

SiPearl is working in close collaboration with its 27 partners from the European Processor Initiative (EPI) consortium - leading names from the scientific community, supercomputing centres and industry - which are its stakeholders, future clients and end-users.

SiPearl employs 130 people in France (Maisons-Laffitte, Grenoble, Massy, Sophia Antipolis), Germany (Duisburg) and Spain (Barcelona).

Media contact:

Marie-Anne Garigue / Grégory Bosson
+ 33 6 09 05 87 80 / + 33 6 60 75 71 61
marie-anne.garigue@sipearl.com
gregory.bosson@sipearl.com

Funded by the European Union

European Processor Initiative