Maplurinum — One Machine out of Many
or
We had 64-bit, yes. What about second 64-bit?

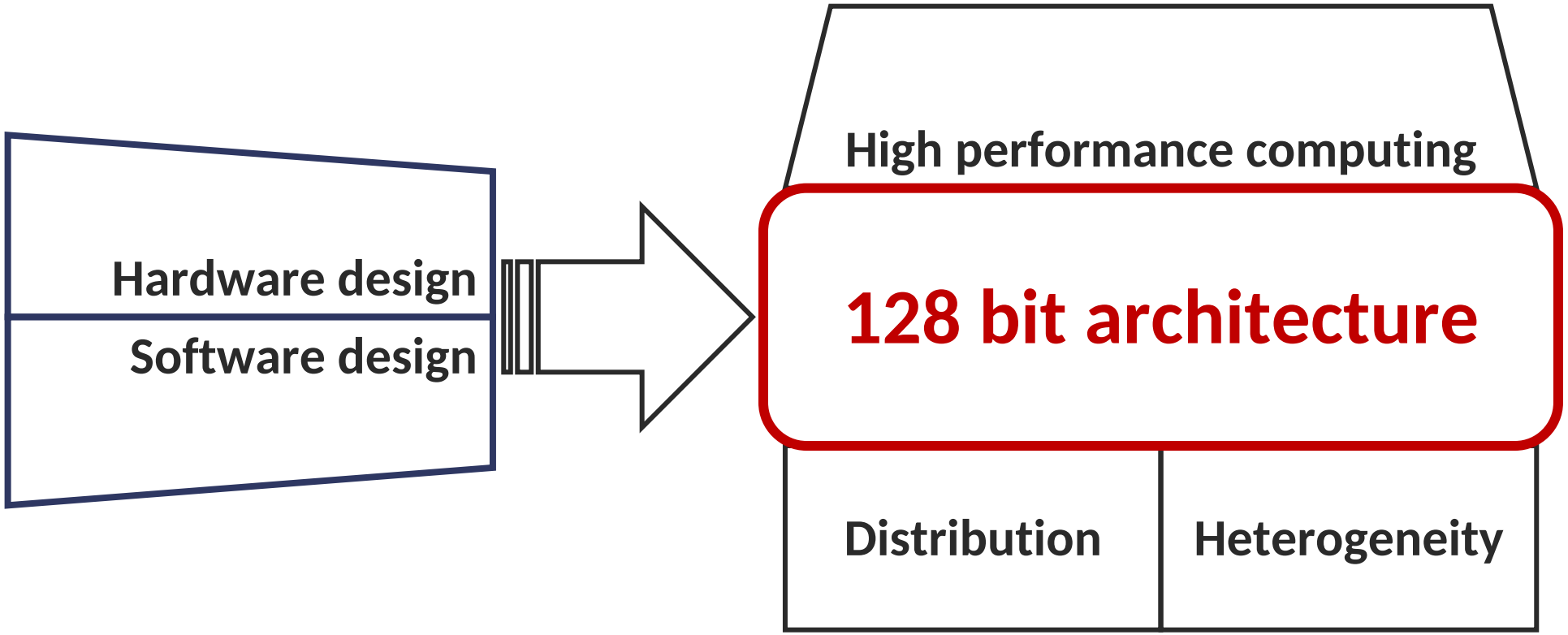
Mathieu Bacou[1], Adam Chader[1], Chandana Deshpande[2], Christian Fabre[3], César Fuguet[3], Pierre Michaud[4], Arthur Perais[2], Frédéric Pétrot[2], Gaël Thomas[5], Jana Toljaga[1], Eduardo Tomasi[2,3]

[1] Samovar, Télécom SudParis, IMT, IP Paris
[2] Université Grenoble Alpes, CNRS, Grenoble INP, TIMA
[3] Université Grenoble Alpes, CEA, List
[4] Inria, Université de Rennes, IRISA
[5] Inria Saclay

French government grant ANR-21-CE25-0016
ANR project « Maplurinum — Machinæ pluribus unum » (Make) one machine out of many

RISC-V Summit Europe, Munich, 27th June 2024

Overview


Hardware design
Software design
High performance computing
128 bit architecture
Distribution
Heterogeneity

HPC TOP 500 — Status & Trends

European machines in the TOP 500 as of November 2023:

- #5 HPE Cray — 2,752,704 cores — Fi
- #6 Bull — 1,824,768 cores — It
- #8 Bull — 680,960 cores — Es

→ **Increasing parallelism and distribution**

Meanwhile:

→ **Trend towards heterogeneity: GPUs, FPGAs, TPUs, variable precision FPUs...**

Hard to use efficiently, hard to program.

Systems	2012 BG/Q Computer	2022	Difference Today & 2022
System peak	20 Pflop/s	1 Eflop/s	O(100)
Power	8.6 MW (2 Gflops/W)	~20 MW (50 Gflops/W)	
System memory	1.6 PB (16*96*1024)	32 - 64 PB	O(10)
Node performance	205 GF/s (16*1.6GHz*8)	1.2 or 15TF/s	O(10) - O(100)
Node memory BW	42.6 GB/s	2 - 4TB/s	O(1000)
Node concurrency	64 Threads	O(1k) or 10k	O(100) - O(1000)
Total Node Interconnect BW	20 GB/s	200-400GB/s	O(10)
System size (nodes)	98,304 (96*1024)	O(100,000) or O(1M)	O(100) - O(1000)
Total concurrency	5.97 M	O(billion)	O(1,000)
MTTI	4 days	O(<1 day)	- O(10)

Source: J. Dongara, Grenoble Sep. 2019.

Big thanks to Henri-Pierre Charles (CEA).

A RISC-V HPC machine by 2030: vision and rationale

At historic rates of growth, it is possible that greater than 64 bits of address space might be required before 2030.
Let's assume that a full RISC-V 128 bit HPC machine could have (wild guess) 100 x 10^6 cores, as 1,000,000 heterogeneous clusters of 100 cores each with o(10 TB) RAM/cluster.

The challenge is how to take advantage of RISC-V and 128 bit to

- Manage the heterogeneity of the machine
- Optimize and simplify the operating system stack
- Increase the performance in distributed computing

Do not beat around the bush:
flat 128-bit address spaces will be adopted
as the simplest and best solution.

> "*There is only one mistake that can be made in computer design that is difficult to recover from — not having enough address bits for memory addressing and memory management.*"
>
> Bell and Strecker, ISCA-3, 1976.

RV128 spec is not frozen at this time, as there might be need to evolve the design based on actual usage of 128-bit address spaces.

Opportunities and challenges for RISC-V and 128 bit

Architecture and Microarchitecture

What could and should change with 128 bit addresses?

1. Allow for disaggregated hardware as a single system image using RV128 as a common denominator
2. RV128 does not mean fully 128 bit microarchitecture

Operating System & Software

What would a 128 bit OS look like?

1. Redesign the process abstraction using hardware-assisted virtualization to bypass the kernel
2. Access to the hardware goes through a unified 128 bit address space

Opportunities and challenges for RISC-V and 128 bit

Architecture and Microarchitecture

What could and should change with 128 bit addresses?

1. Allow for disaggregated hardware as a single system image using RV128 as a common denominator
2. RV128 does not mean fully 128 bit microarchitecture

Operating System & Software

What would a 128 bit OS look like?

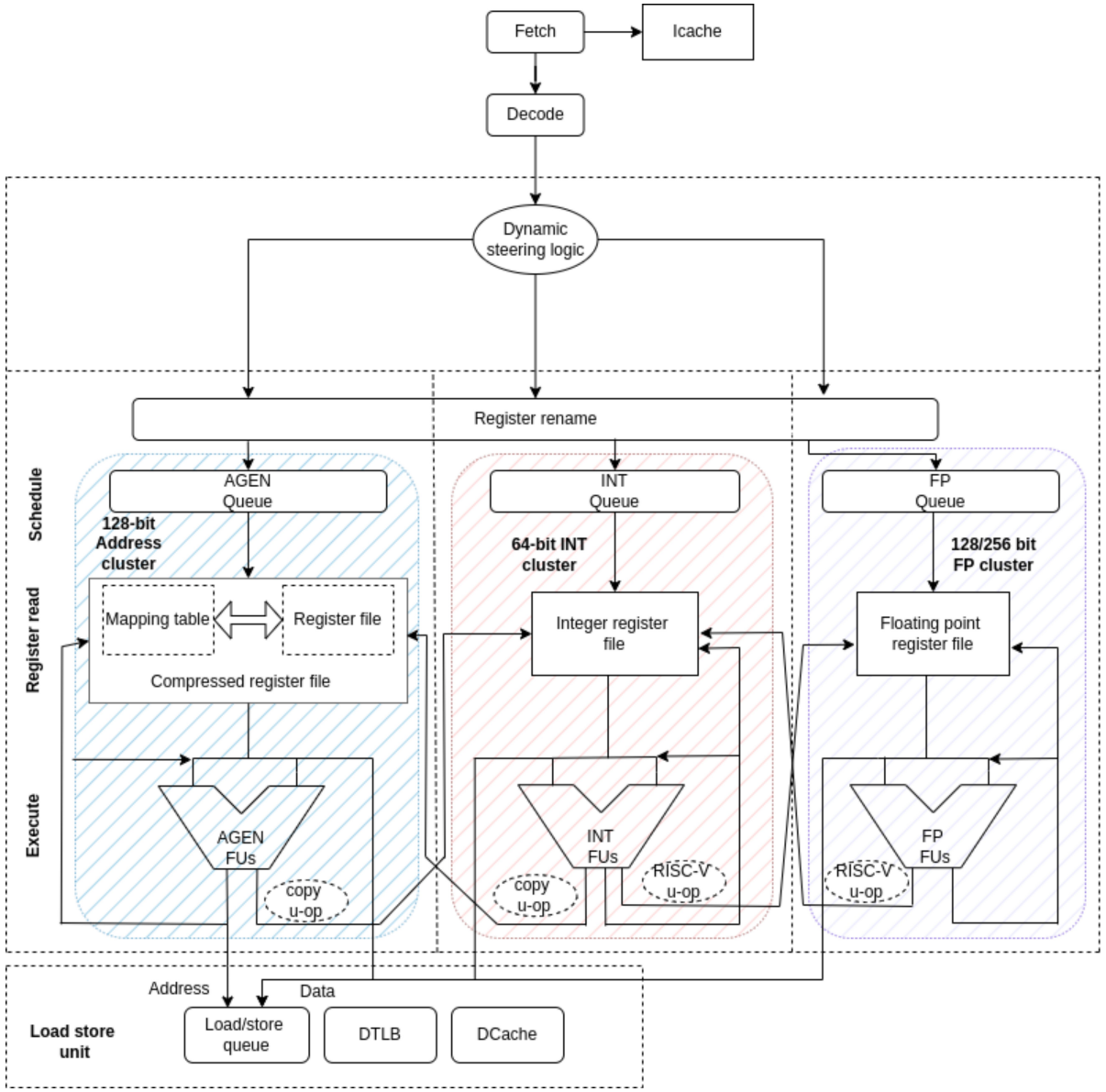
1. Redesign the process abstraction using hardware-assisted virtualization to bypass the kernel
2. Access to the hardware goes through a unified 128 bit address space

Opportunities and challenges: microarchitecture

- Most 128-bit operations will (likely) manipulate addresses

- Opportunity for clustered microarchitecture
  - ADDR (128b) vs. INT (64b) vs. FP/SIMD (128/256/512...)
  - ISA-agnostic (hardware steering)

- Also
  - 128-bit pointers likely very compressible -> memory layout to favor cache compression (put all your pointers together)



Opportunities and challenges for RISC-V and 128 bit

Architecture and Microarchitecture

What could and should change with 128 bit addresses?

1. Allow for disaggregated hardware as a single system image using RV128 as a common denominator
2. RV128 does not mean fully 128 bit microarchitecture

Operating System & Software

What would a 128 bit OS look like?

1. Redesign the process abstraction using hardware-assisted virtualization to bypass the kernel
2. Access to the hardware goes through a unified 128 bit address space

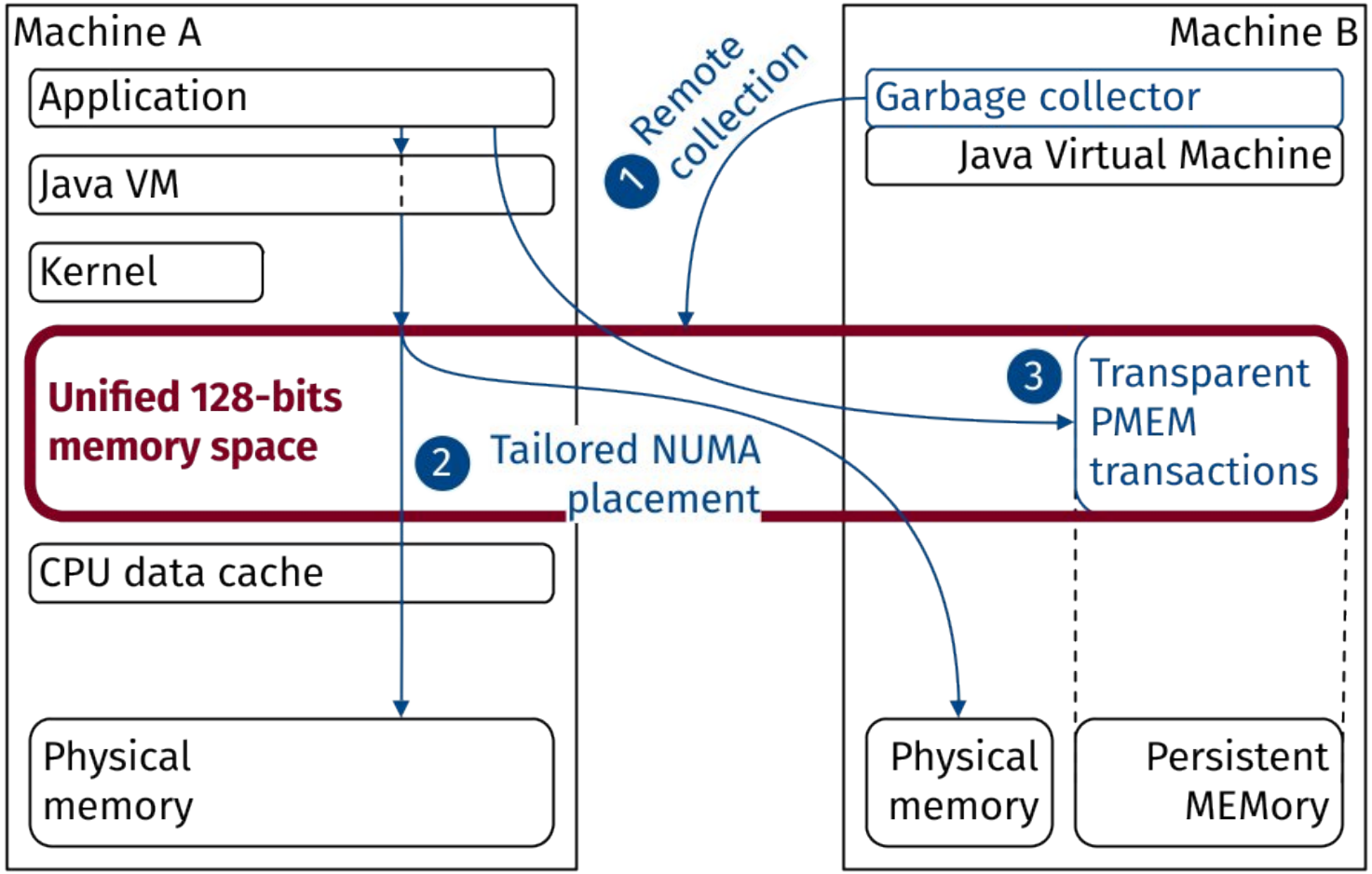
Opportunities and challenges: OS & software

The operating system is just a controller that grants access to the hardware.

- Common interface: unified 128 bit memory space via hardware-assisted virtualization

Many uses:

1. Disaggregated runtimes (e.g. Java VM)
   - Remote garbage collector to prevent cache pollution at the application side
2. NUMA placement piloted by the app
3. Direct transparent use of Persistent Memory via transactions based on hardware memory management
4. ...



Conclusion

1. RISC-V 128 bit is an opportunity for hardware – software co-design
2. The opportunities and real issues are in parallel and distributed aspects of future RISC-V 128 bit machines, not so much in *classical* ISA extensions
    1. Disaggregated OS over a unified address space
    2. Common general purpose 128 bit architecture over heterogeneous hardware

- Though 128 bit machines are probably far away, such work will take time, so we are starting now!

Questions?

Please come see our poster today at D-04

Work funded by the project
« **Maplurinum — Machinæ pluribus unum** »
(Faire) une seule machine avec plusieurs
(Make) one machine out of many
French gov. grant n° ANR-21-CE25-0016