We had 64-bit, yes. What about second 64-bit?

Mathieu Bacou[⋆], Adam Chader[⋆], Chandana Deshpande[§]
Christian Fabre[⋆⋆], César Fuguet[⋆⋆], Pierre Michaud[†], Arthur Perais[§],
Frédéric Pétrot[§], Gaël Thomas[‡], Jana Toljaga[⋆], Eduardo Tomasi[⋆⋆§]

[⋆]Samovar, Télécom SudParis, IMT, IP Paris [⋆⋆]Univ. Grenoble Alpes, CEA, List
[†]Inria, Univ. Rennes, CNRS, IRISA [‡]Inria Saclay
[§]Univ. Grenoble Alpes, CNRS, Grenoble INP[1], TIMA
[1]Institute of Engineering Univ. Grenoble Alpes

Overview of ANR Project Maplurinum (ANR-21-CE25-0016)

The Big Picture

- Current OSes struggle to provide efficient abstractions for increasingly heterogeneous hardware (accelerators)
- Scalability issues of hardware and software to a rack-scale computing model where multiple blades share main memory
- Advent of load/store accessible backup store (e.g., NVM)

The Goal

- Rethink the operating system with a disaggregated userspace controlling hardware features
- Future-proof it by implementing 128-bit flat addressing
  → Rely on an open-source RISC-V extension, RV128
  → Impacts the whole stack from OS to micro-architecture

Operating System & Software

Disaggregated userspace

- Scalability and uniformity over hardware heterogeneity
  → Features for the disaggregated userspace runtimes
  → Kernel bypassing for scalability using virtualization

Unified address space

- Allow direct access of applications to the data plane: loads and stores to unified memory space
  → User processes manage their own virtual memory space
  → User processes control their disaggregated memory space

Architecture & Microarchitecture

128-bit Architecture

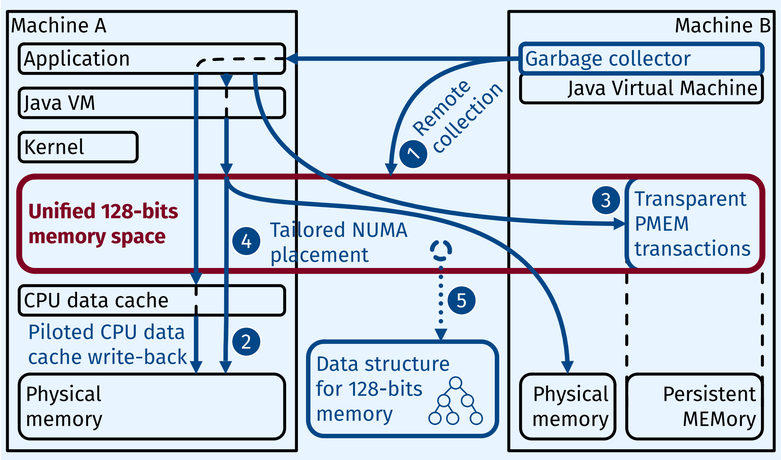
- RV128 extension as a common denominator: All agents (CPUs, GPUs, TPUs, FPGAs) are RV128-capable
  → Disaggregated software can run anywhere

128-bit General Purpose Microarchitecture

- Naively: Double datapath width (bypass, registers, functional units)
- Dennard and Moore scaling not there to absorb the change anymore: Would like to limit hardware cost of RV128

Machinæ pluribus unum – One Machine out of Many

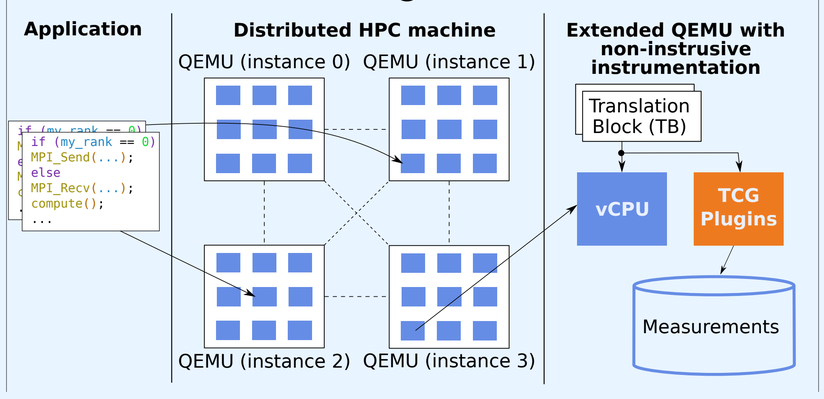
Operating System



- Efficient memory disaggregation: user-controlled caching
  1. Java VM with a remote Garbage Collector to prevent cache pollution
  2. Piloted CPU data cache for sync.-free GC
- Unified address space: hardware virtualization for the userspace
  3. Transparent persistent memory transactions
  4. Tailored NUMA placement
  5. Data struct. optimized for 128-bits memory

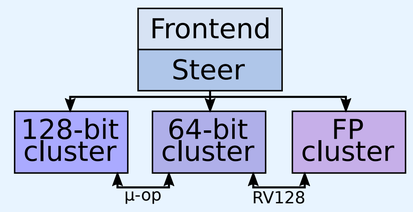
Memory Hierarchy

- NUMA effects are exacerbated in heterogeneous multi-socket, multi-board HPC computing systems.
- 128-bits shall ease the programming of large scale systems, but NUMA effects must be considered.
- Software and hardware mechanisms are being analyzed to hide this latency.
- We developed a QEMU-based simulator for distributed large scale machines.



Microarchitecture

- Compile an existing C program to RV128: About 40% of the instructions still operate on 32/64-bit
  → 128-bit operations will mostly be address generation slices



- Divide & Conquer: 128-bit cluster for addresses, 64-bit cluster for arithmetic
  → Push complex 128-bit operators (e.g., mul, div) to SW
  → Compress addresses (PRF, TLBs tags/data, cache tags), reduces area
  → Introduce address type in ISA?

Perspectives

- Discover the minimal ISA and adequate software interfaces for the disaggregated userspace
- Establish the user-kernel interfaces for efficient distributed computing through the unified address space
- Revisit basic operating system concepts for machine-wide, unified 128 bit address space: process, memory mapping policies, etc.
- Identify hardware requirements for adequate support of a machine-wide, distributed 128 bit address space
- Propose OS-driven hardware mechanisms that replicate data between nodes and manage coherency to reduce NUMA latency