Extended Main System Memory over OpenCAPI

Dimitris Syrivelis, IBM Research - Ireland
Overview

• "Stealing" Main System Memory from Remote Servers
Why?

- Dynamic composition of systems with very large main system memory (even if slower than local):
  - Appropriate for in-memory data grids (Memcached, SAP-Hana, Redis etc..)
  - Instead of scaling-out in software for memory we scale-up systems in hardware
Why?

• Dynamic composition of systems with very large main system memory (even if slower than local):
  • Appropriate for in-memory data grids (Memcached, SAP-Hana, Redis etc.)
  • Instead of scaling-out in software for memory we scale-up systems in hardware

• Better utilization of resources under public cloud workloads:
  • Resources tend to get ”jailed” on servers (i.e. cores available but no memory and vice versa)
Why?

• Dynamic composition of systems with very large main system memory (even if slower than local):
  • Appropriate for in-memory data grids (Memcached, SAP-Hana, Redis etc.)
  • Instead of scaling-out in software for memory we scale-up systems in hardware

• Better utilization of resources under public cloud workloads:
  • Resources tend to get ”jailed” on servers (i.e. cores available but no memory and vice versa)

• ..because with OpenCAPI we can! And we want to explore this radical disaggregated systems avenue.
Key Components Overview: Memory Host Agent

- PCI-scan Info
- Extended System Memory Owner
- MMIO Registers
- OpenCAPI Transceiver Clock Domain (400Mhz)
- Network Interface Clock Domain
- Memory Request Stream
- Request Response Stream
- Network Dispatch
- Network Receive
- Fifos
- Fifos
- Network Rx/Tx
- Error Report Interrupt
- Can Map up to 4TB Physical Address Space Per OCAPI Brick

IBM Research
Key Components Overview: Memory Host Agent

- Extended System Memory Owner
- PCI-scan Info
- MMIO Registers
- Memory Request Stream
- Request Response Stream
- Error Report Interrupt
- OpenCAPI Transceiver Clock Domain (400Mhz)
- Network Dispatch
- Network Receive
- Fifos
- Local Caching Support to be included

Can Map up to 4TB Physical Address Space Per OCAPI Brickx
Key Components Overview: Memory Provider

Network Interface Clock Domain

- Network Rx/Tx
- Fifos
- Network Dispatch
- Network Receive

OpenCAPI Transceiver Clock Domain (400Mhz)

- Memory Request/Response Handling Engine
- PASID, Effective Address Base
- MMIO Registers

PCI-scan Info

Memory Provider
Key Components Overview: Software

- Host Memory Controller
- Linux Kernel New Numa memory node HotPlug
- Skiboot Physical Address Space BAR Dynamic Config
- Extended System Memory Owner
- Rack Memory Resource Manager
- System Memory Allocator Agent
- LibOCXL

System Memory Providers
Key Components Overview: Software

Host Memory Controller

Linux Kernel New Numa memory node HotPlug

Skiboot Physical Address Space BAR Dynamic Config

Extended System Memory Owner

LibOCXL

System Memory Allocator Agent

System Memory Providers