Rack-scale Disaggregated Cloud Datacenters & the dReDBox project

MULTIPROG Workshop 2017

Dionisios Pnevmatikatos
FORTH
Valentina Salapura (IBM):
- IBM is building disaggregated servers for cloud

Babak Falsafi (EPFL):
- Build around the memory
- Use aggregate memory over the network (mem. caching services)
- Rack-scale with PCI+TCP/IP ~ few usecs
- Specialize servers around DRAM

This Conference:
- Data center efficiency & acceleration in the frontline!
Motivation

- Memory-centric data-center use
- Network end-to-end latency decreasing
- dReDBox ambition: increase resource disaggregation
  - Focus on system integration
  - Compute/memory/accelerator HW blocks
  - Flexible combination of above for node customization
  - Deep SW-defined control of resource allocation

A customizable low-power datacenter architecture
Shift from the mainboard-as-a-unit paradigm to a flexible software-defined block-as-a-unit
dReDBox project & consortium

- H2020 research project #687632
- 11 partners
- 7 countries
  - Ireland (2)
  - Greece (2)
  - UK (2)
  - Spain (3)
  - Netherlands
  - France

Dionisios Pnevmatikatos, January 24, 2017
Introduction: Disaggregation roadmap
Key ideals/features

- Three main building HW blocks (SoC, memory, accelerators/FPGA)
- Mix-and-match node configuration of building blocks
- Novel scalable **optical network** interconnects memory controllers / modules in a datacenter fashion
  - Multi Tbps switch bisection, software-controlled, fixed minimum latency (switch terminal I/O to switch terminal I/O)
  - Serialized memory interface (between nodes)
- Hypervisor support for resource allocation & VM execution
  - High-level resource provisioning and management
  - IPMI (v2) resource management for reduced power consumption
dReDBox architecture

- Server board (Tray)

Tray with hot-pluggable and interchangeable bricks (CPU, RAM, Accelerators)
dReDBox Memory disaggregation

- Local memory module(s) used for bootstrapping/startup
- dReDBox will introduce dedicated memory blocks, interfaced to the microserver (SoC) blocks
- Challenge: Develop an appropriate interface / logic for transmission over the optical network

Virtual Memory Modules
- Local memory interfaced directly to the memory controller
- Memory pool blocks accessible by all physical modules
- Partitioning of physical memory (controlled by system management SW)
- Goal: RDMA ~= DMA
dReDBox architecture (cntd)

- **OS support for disaggregation**
  - KVM based hypervisor (HV)
  - One HV instance will run on each microserver brick
  - Microserver BIOS will identify only locally attached components
  - Disaggregated memory dynamic allocation at Orchestration Tool level
Resource allocation & orchestration

- Challenges:
  - Datacenter-wide Global Address Space
  - IPMIv2 control per component

- New orchestration tools
  - Dynamic platform synthesis
  - HV support for memory segmentation and ballooning
  - Power management w/ IPMIv2
dReDBox accelerators

RC primarily, but not exclusive!

- Defined accelerator module(s)
  - Xilinx UltraScale+ for prototyping

- Option #1
  - Exploit the ultra-low latency, dedicated links with compute bricks

- Option #2
  - Use a disaggregated, over-the-network I/F

- Ongoing, hot topic!
Work layers

Use cases

Orchestration

SW

Hypervisor/
System
Software

Arch, HW
Prototypes & Optical Network

Telefonica

naudit

Kinesense

IBM

UCL

UNIVERSITY OF THESSALY

Virtual Open Systems

UNIVERSITY OF THESSALY

FORTH

BSC

polatis

UNIVERSITY OF THESSALY

sintecs

UCL
Thank you!

Questions?