Communication In Massively Parallel Cloud Computer Systems

Alan Mujumdar and Simon Moore

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Motivation

Over the past few years the gap between CPU performance and I/O bandwidth has continually grown. Bottlenecks due to storage, network configuration and memory bandwidth create low CPU utilization. Networking elements such as routers, switches and network interface cards, contribute towards datacentre power consumption. Large datacentre networks suffer security issues. A robust communication scheme within datacentres will help combat vulnerabilities. Majority of failures and the resulting down time occur due to the network.

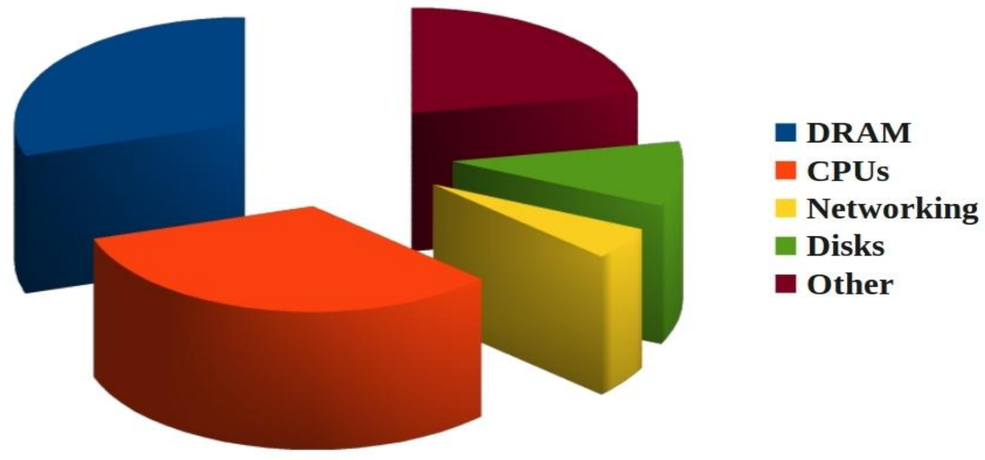

Figure 1
Power consumption in Google datacentres (2007)

Research

The goal is to reduce power consumption, latency and improve security in datacentres. Data gathered in Google datacentres shows that the network consumes 5% of total power and this is likely to increase (Figure 1). In other datacentres this figure could be as high as 10% to 20%. Cutting down on network elements could reduce these costs. The network is also a major latency contributor. Nodes in large symmetric multiprocessors experience communication delays in the orders of 100ns, LAN based systems suffer latencies of 100us and above. Introducing on-chip routers could reduce these numbers.

Intended Outcome

A typical datacentre setup is shown in Figure 2. It is clear that switches are a bottleneck. The proposed design is illustrated in Figure 3. Due to the absence of low level switches/routers, some barriers are eliminated. In order to test this concept, a multi-core soft processor system will be designed. Currently all research is based on a Altera DE4 FPGA board. A finalised version of the system will look similar to Figure 4. The setup will be used to analyse cache coherency schemes and verify the feasibility of a physically distributed memory model.

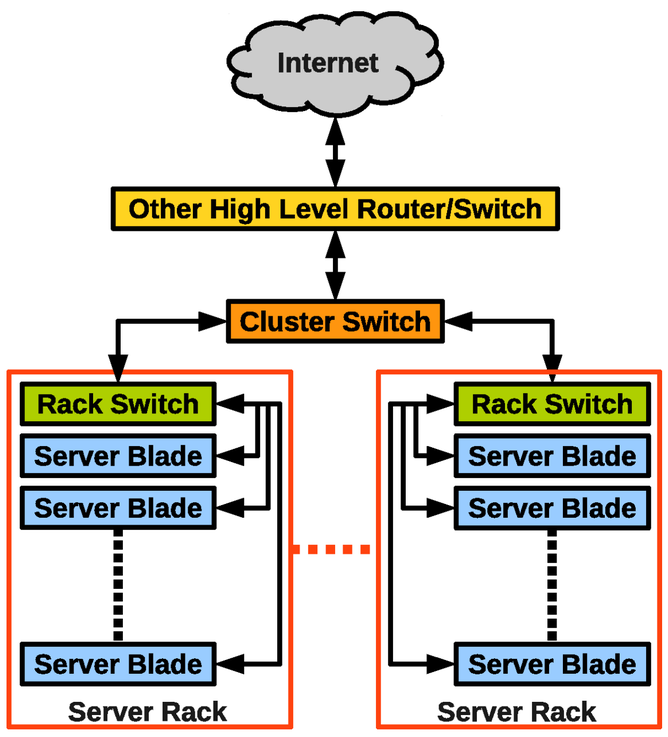

Figure 2
Typical datacentre setup

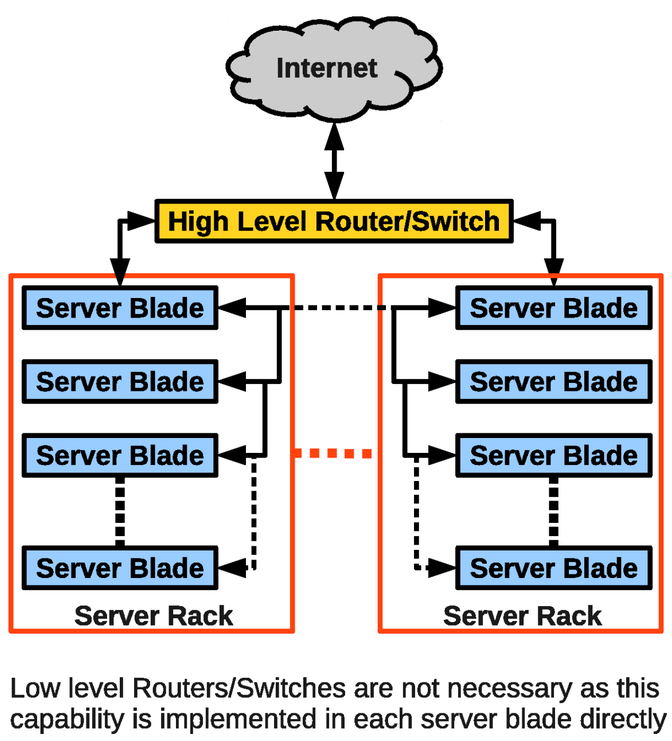

Figure 3
A datacentre based on the proposed architecture

Current Research

A dual-core version of the CHERI processor is currently being designed (Based on the BERI platform). This version uses a shared L2 cache and the MESI algorithm.

Future Work

The CHERI processor design will be extended to incorporate a variable number of cores, different levels of cache and cache coherency algorithms. The design will allow a single multi-core CHERI processor to be split across several FPGA boards. This will check the validity of a physically distributed cache. Multiple network configurations will be implemented in order to determine the practicality of this design.

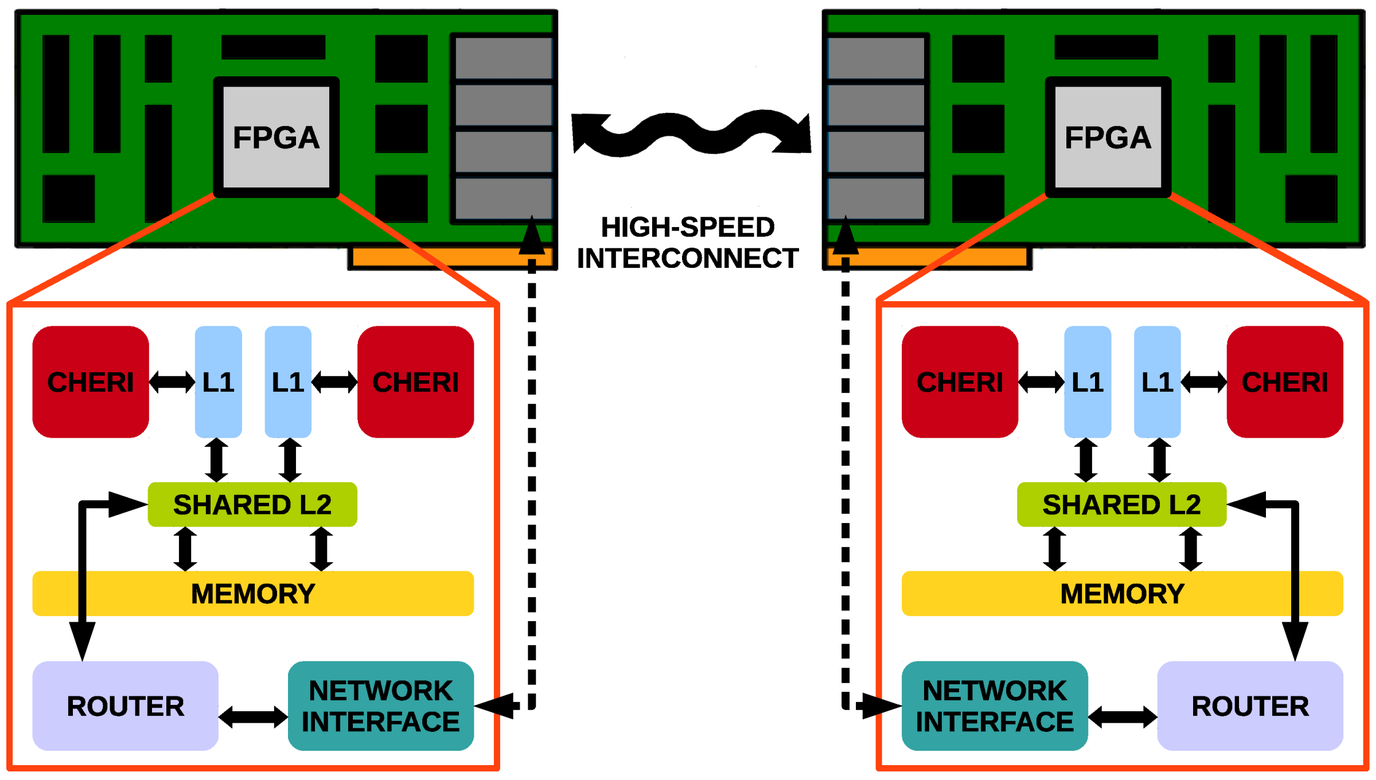

Figure 4
Test hardware setup

BERI platform funded by DARPA

Approved for public release. This research is sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL), under contract FA8750-10-C-0237. The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

alan.mujumdar@cl.cam.ac.uk simon.moore@cl.cam.ac.uk

Computer Architecture Group
http://www.cl.cam.ac.uk/research/comparch/