CMP SCI 635: Modern Computer Architecture

Interconnections in Multi-core Architectures: Understanding Mechanisms, Overheads and Scaling

Name: Kunjal Panchal Date: 14th Nov, 2019 Student ID: 32126469

Paper: 19 – Parallelism [Kumar05]

Strengths:

1. In integrated circuits, the high density of wires has increased coupling problems and further scaling down of wire size can actually slow down the circuit because of the increased resistance. These problems have been mitigated by manipulating the signal using serial links and wire-aware designs. However, the core issues still remain. Which this paper addresses well.
2. The paper proves that omitting large coherent caches, although diverging from the well-known shared memory model that many scientific applications follow, enables the design of less complex and fast memory. The area reclaimed from the omission of cache and snooping control logic can be used for better connectivity features of multicore CMPs.

Weaknesses:

1. Traditionally, the Network on Chip has been designed, implemented and evaluated assuming electronic routers and metal wires. Maybe at the time this paper was written, there were no non-wire alternatives. But, can't we consider, say, an optical interconnection network integrated on a chip? Is this connecting cores on a same chip problem even relevant now?
2. "*We do not assume off-chip L3 cache, but in 65nm systems, it is likely that L3 chips would be present as well (the number of L3 chips would be limited, however, due to the large number of pins that every L3 chip would require), but we account for that effect using somewhat optimistic estimates for effective bandwidth and memory latency.*"
   But as the size of L3 cache increases, the increase in pincount won't be proportional (it will be logarithmic), so even if number of L3 chips don't increase but the capacity increases, the memory latency will be significantly affected.

Questions/Assertions:

1. Has any of you deployed Intel OmniPath on a large scale? (talking about 100+ nodes)? If yes, how can it be compared to the experience with InfiniBand?

2. Why AMD Zen is using 2 CCX modules? As far as I know, not for yields because it's still one die so yields shouldn't be better. Why not make just one 8core CCX? No CCX to CCX latency problem, 1x16MB L3 cache available for each core.

3. GPU and CPU interconnects are entirely different ball games. One is high bandwidth as a priority and another is low latency, all the while operating at frequencies that are 2-4 GHz apart. How does hardware architects tackle this?

4. Is there any reason why they could not run the Infinity Fabric faster? 3.2 GHz was possible on HyperTransport from which Infinity Fabric was based on. Right now, by default, with DDR4-2133, it's at 1/3 that at just 1066 MHz. Judging by the results, most of the problems are resolved with DDR4 3200 with tight timings, suggesting that even a 50% improvement to 1600 MHz could solve most of the worst problems. Doubling the fabric speed to say, 2133 MHz for the fabric ought to address many of the latency issues. What would the power consumption penalty be?

5. Intel hinted towards a Xe 'Coherent Multi-GPU' future with CXL Interconnect. Is it any different than Nvidia NVLink?

6. DX12 enables developers to utilise multi GPU scaling in ways never possible before. They can mix AMD and Nvidia cards, they can use cards for different parts of the rendering pipeline and all sorts of fancy stuff. But why there is not much interest for multi GPU scaling?