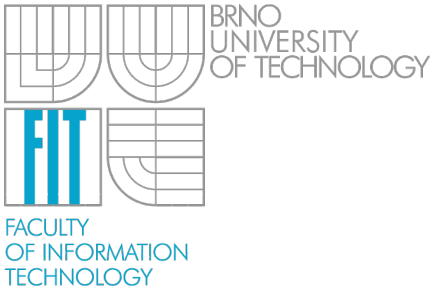
100G Ethernet a výhled do budoucna

Research Group

Jan Kořenek

Brno University of Technology, Faculty of Information Technology
Bozetechova 2, 612 00 Brno, CZ
http://merlin.fit.vutbr.cz/ant/



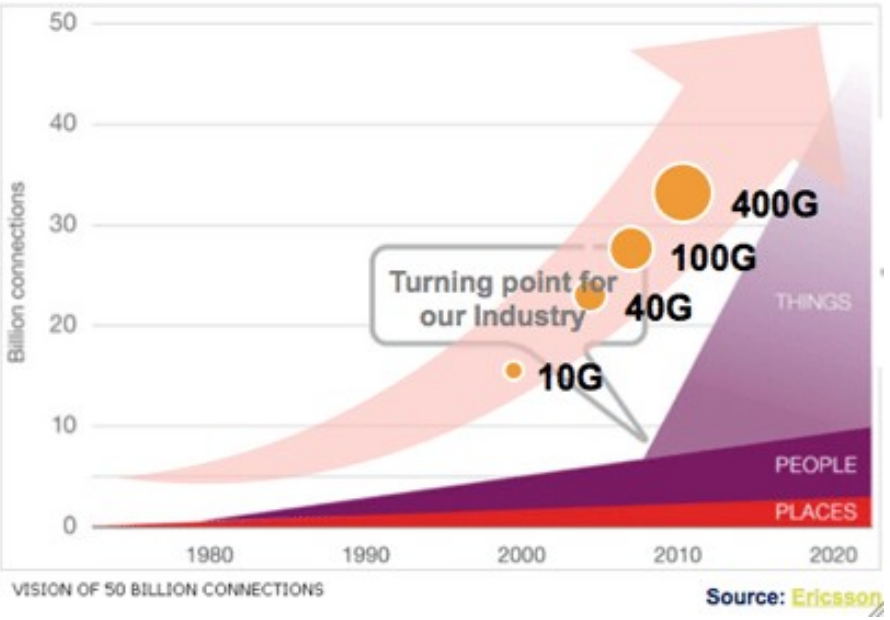
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Osnova

- Proč 100Gb Ethernet
- Omezení pro návrh 100Gb zařízení
- Datová centra a switche

Motivace

- Předpoklad více než 50 miliard připojených zařízení do konce roku 2020
  - ***Rozšíření videa a grafiky***
  - ***Sdílení na sociálních sítích***
- Rozšíření 100Gb sítí zejména na páteřních linkách a v DC



- Základním problém zařízení pro 100Gb je paměť
  - Omezená rychlost I/O může snižovat propustnost
  - Dlouhý čtecí cyklus pamětí zpomaluje zpracování paketů
  - Vysoká spotřeba může přesahovat možnosti napájení

Datová cesta

- Propustnost bufferu pro ukládání paketů:
  - 100Gb pro zápis paket + 100Gb pro čtení = celkem 200Gb
  - U běžných DDR3 pamětí je ale režie komunikace až 50%
  - Pro buffrování paketů je nutné mít paměť s propustností 300Gb
  - Propustnost jedné DDR3-1600 je cca 25,6 Gb
  - Pro realizaci 100Gb řešení je potřeba 12 DDR3 pamětí (12Gb paměti)
  - Pakety délky 65B bajtů mohou způsobovat další snížení .propustnosti, režii
  - Většinou jsou adresové vodiče DRAM sdíleny tak, aby více modulů tvořilo 64b datovou sběrnici. Pro připojení pamětí je potřeba **400 pinů**
- Kapacita bufferu pro ukládání paketů ve směrovači / switchi
  - Ve směrovačích/switchi je na páteřní síti většinou buffer na 5 až 10 ms
  - Kapacita potřebná pro uchování paketů 0,5Gb
  - Pro koncové zařízen (RTT=2x100ms) je velikost bufferu kolem 20Gb

Řídící cesta

- Propustnost paměti s řídicími informacemi:
  - Pro dosažení 100Gb je nutné zpracovat 150 Mpkt/s
  - V průměru 14 náhodných přístupů do paměti na jediný paket (malá informace)
  - Reálně 2Giga přístupy za sekundu – je nutné použit RLDRAM nebo QDR SRAM
  - RLDRAM nebo QDR SRAM vyžadují cca 600 I/O pinů

Požadavky na I/O piny procesoru

- Datová cesta (DDR3 DRAM) …. 400 pinů
- Řídící cesta (RL DRAM nebo QDR SRAM) …. 600 pinů
- Celkové požadavky na datovou a řídící cestu …. ~ 1000 pinů

- Vysokorychlostní sériové kanály snižují požadavky na I/O
  - Datová cesta … 128 pinů
  - Řídící cesta … 64 pinů

- Rozhraní MASI – Multi Access/Serial Interface (MASI)

Požadavky na čtecí cyklus tRC

- Zpracování paketů na vstupu i výstupu (celkem 200Gb)
- Na 100 Gb síti může přijít až 150 Mpkt/s (6,67 ns)
- Vstup i výstup - čtecí cyklus 3,33 ns
- Přístupy pro snížení tRC:
  - Vyšší frekvence jádra paměti – extrémní nárůst spotřeby zejména díky parazitním kapacitám
  - Replikace dat do více paměťových banků (multibank strategy) – dvě kopie tabulky sníží tRC na polovinu
- Současné DDR3 paměti mají tRC=46ns – pro 100Gb ~14 kopií
  - Prakticky nepoužitelné díky ceně, spotřebě, složitosti desky, …
- Kompromis mezi vyšší frekvencí a replikací jsou MASI paměti

Power Budget

- Nárůst propustnosti – vyšší počet 10Gb portů nebo 100Gb, cena
- Spotřeba je klíčovým návrhem současných síťových systémů:
  - „In 100G design centers, cost is no longer measured in dollars or ms of latency, but in Watts.“
  - Pětkrát prodražuje řešení: napájecí zdroje, chlazení na desce nebo v RACKu, klimatizace v místnosti, spotřeba záložních zdrojů, spotřeba power managementu
- Pro snížení ceny je nutné dosáhnout nižší spotřeby na vyšší komunikační rychlosti
- Omezení návrháře z pohledu power budgetu na desce, teplotních omezení, atd.
- Jednou z možností redukce spotřeby je pomocí MASI paměti (~2,5 x)

Interface	ASIC Pins	Throughput	Power	Gbps/W
DDR-3 x 16	144	26 Gbps	1.5W	17.33
12 (DDR-3 x 16)	1728	312 Gbps	18W	17.33
MASI x 8	66	72 Gbps	1.7W	42.35
4 MASI x 8	264	288 Gbps	6.8W	42.35

Figure 2: Comparison between typical parallel and serial device-to-device memory (MASI) configurations. Notes:
1) "ASIC pins" include all power & grounds
2) DDR3 at 1600 Mts, GCI(MASI) at 10 Gbps on 16 lanes
3) 12(DDR-3 x16) do not share addr/cmd
4) Throughput = Read + write bandwidth

Datová centra a switche

- Přechází se z 1Gb na 10Gb – Google a Facebook by chtěl přejít na 1T core v roce 2013 :-)
- Klíčová je nízká latence switchů, z pohledu odezvy
- Změna návrhu switchů – přechod od crosbaru k pamětím.

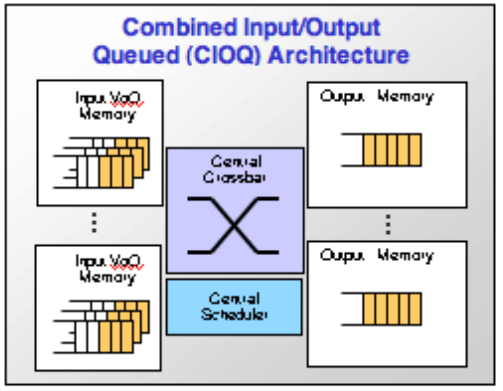


Figure 1. Combined input/output queued (CIOQ) architecture.

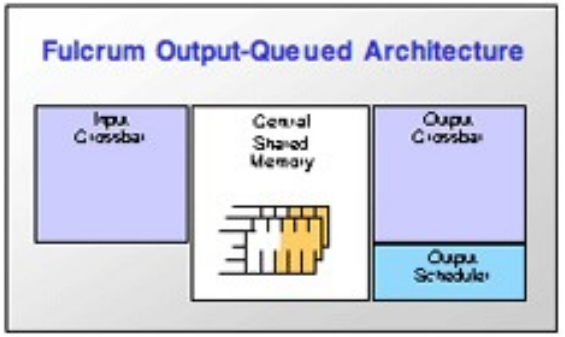


Figure 2. Output queued, shared memory architecture with extremely low latency.