
Tecnologie per high speed link (e dintorni):
(pochi) fatti e (molte) ipotesi

Una visione molto personale
e sicuramente parziale...

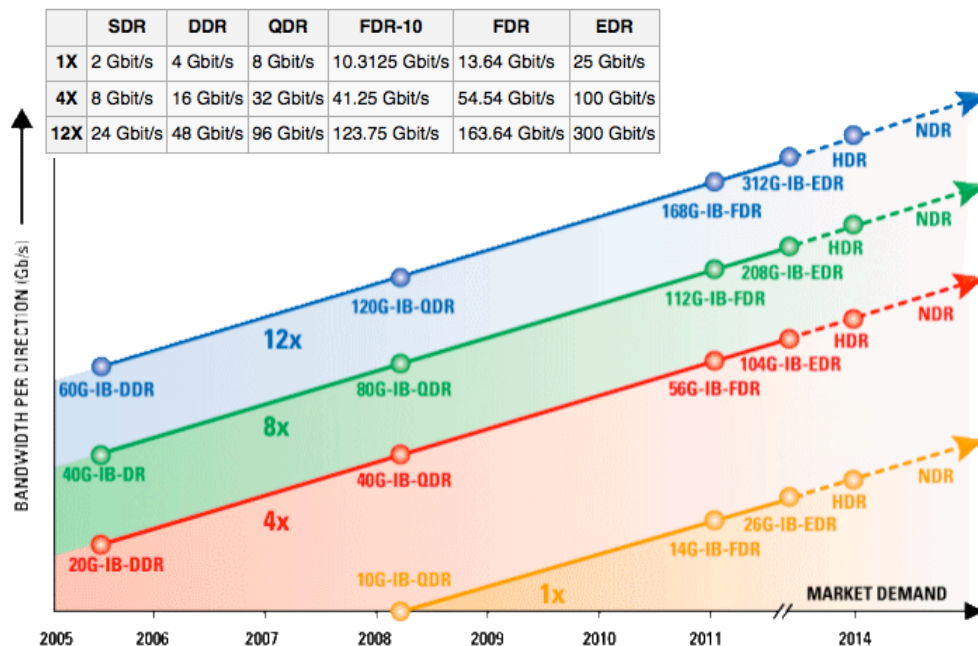
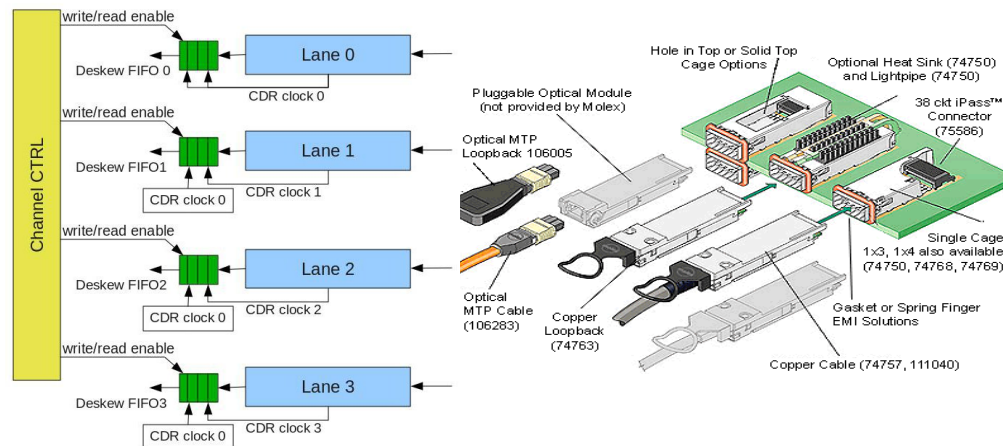
P.V. & A.B.

GAP general meeting – Roma - 10/5/2013

- Se si parla di tecnologie di link... magari applicate a sistemi di calcolo paralleli (in qualche senso) customizzati, GPU-based e dedicati ad esperimenti HEP e/o medical imaging bisogna considerare:
 - Roadmap tecnologica per physical layer di **serial link**
 - Utilizzo in **canali di trasmissione**
 - Integrazione in **FPGA**
 - Roadmap per **GPU**

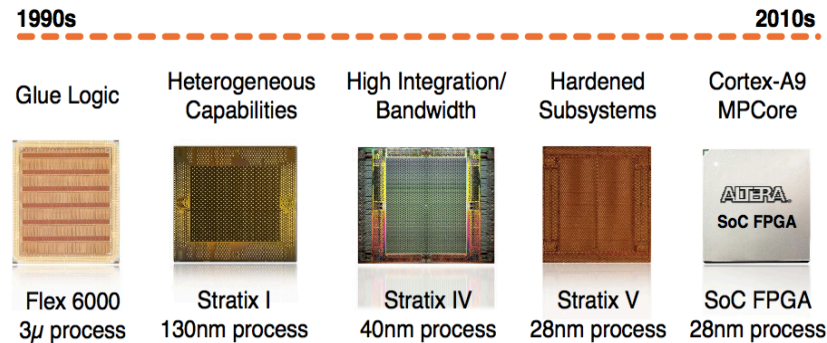
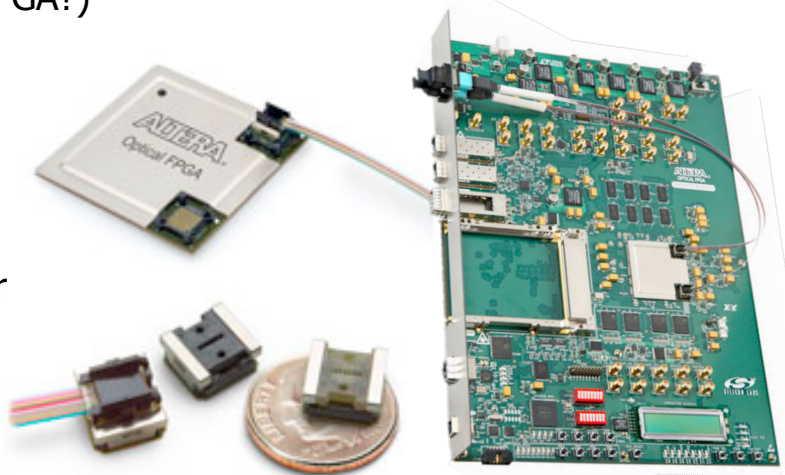
- Serial link
 - Oggi **8-10 Gbps** per linea
 - Tecnologia differenziale, **CML**, power hungry...
 - **CDR** (Clock Data Recovery):
 - Completo decoupling tra sender e receiver
 - PLLs, protocolli di sincronizzazione complessi, meccanismi sofisticati di error check,...
 - Encoding per link robustness
 - **8b/10b** (20% overhead) -> **128b/130b**
 - **Copper passivo** (<10mt), **attivo** (<20m), **Ottico**
 - Chip dedicati (serdes, protocol encoder+phy,...)
 - Embedded links in FPGA high end

- Performance del singolo link
 - Oggi **10Gb/s** -> **25 Gb/s**
 - x2 ogni 3 anni -> **100 GB/s** nel 2020
- Canali di trasmissione realizzati attraverso bonding di serial links + stacks di protocolli superimposed
- Convergenza tecnologica per copper e ottico (almeno a livello della meccanica)
 - **QSFP+**
- **INFINIBAND** come case study
 - Riferimento per canali di trasmissione high-end
 - 2014 EDR 25 Gb/s per linea
 - 2017 -> H(igh)DR e N(ext)DR
 - 50 Gb/s -> 100 Gb/s per linea
- **Ethernet**
 - Roadmap simile (100G -> 300G)
 - Mostra latenze maggiori ma e' piu' economico....
- Problema qui e' la capacita' degli switch di sostenere il traffico integrato di N canali in parallelo
- Soluzione possibile le reti N-Dim torus i.e. **APEnet+** ;)
 - Point-to-point channels
 - Scaling garantito per grandi sistemi
 - Throughput aggregato va linearmente con il numero dei canali



- In generale sono i componenti elettronici piu' complessi ed avanzati. Nel tempo devices sempre piu' evolute che fanno leva sul miglioramento dei processi di produzione
- Roadmap prevede l'uso delle tecnologie state-of-the-art
 - Da puro "sea-of-gates" programmabile... a SoC con dual ARM @800MHz on board (!)
 - Nelle attuali FPGA Tbits di I/O e TFlops di DSP arith blocks integrati
 - Tecnologia stacked 3D nelle prossime device
 - More bandwidth, nuove architetture
- Nelle generazioni correnti
 - 4-10 transceiver @28gbps
 - Nx10 transceiver @10-14 gbps
 - Nx100 transceiver @1-5 gbps
 - Numerose HardIp on board
 - PCIe Gen3/2/1, 10G/40G/100G Ethernet, Serial RapidIO, CPRI...
 - Latenza fissa utile per sincronizzare il timing dell'hardware su link seriali (TTC on FPGA?)

- Link ottici su FPGA
 - Proof of concept AVAGO + Altera
 - Avago MicroPod up to 120gb/s full bidir
 - Waiting for Optical on die

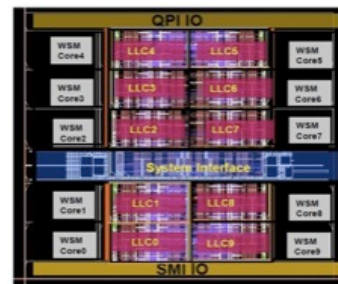


© 2012 Altera Corporation— FPGA 2012 Conference Panel

ALTERA

	ALTERA STRATIX V	XILINX VIRTEX 7
Logic Cell (up to)	1.05M	1.9M
Registers (up to)	1.58M	2.4M
Peak transceiver speed (Gbps)	14.1 28.0	13.0 29.0
Device serial bandwidth (Gbps)	936	930
PCIe interface	up to 4 Gen3,x8	up to 4 Gen3,x8
Memory interfaces	up to 7 banks x72 DDR3@800Mhz	DDR3 1.8Gbps
Embedded Memory size	55 Mb	65 Mb
DSP elementary block	4096 (Dual 18x18 Mul + 64bit Acc)	3960 (25x18 Mul + 48bit Acc)
I/O pins	up to 1125 (+ dedicated transceiver pins)	up to 1200 (+ dedicated transceiver pins)
Hardened IP	NIOS (proprietary) ARM A9 MIPS32 (soft core)	Dual ARM A9

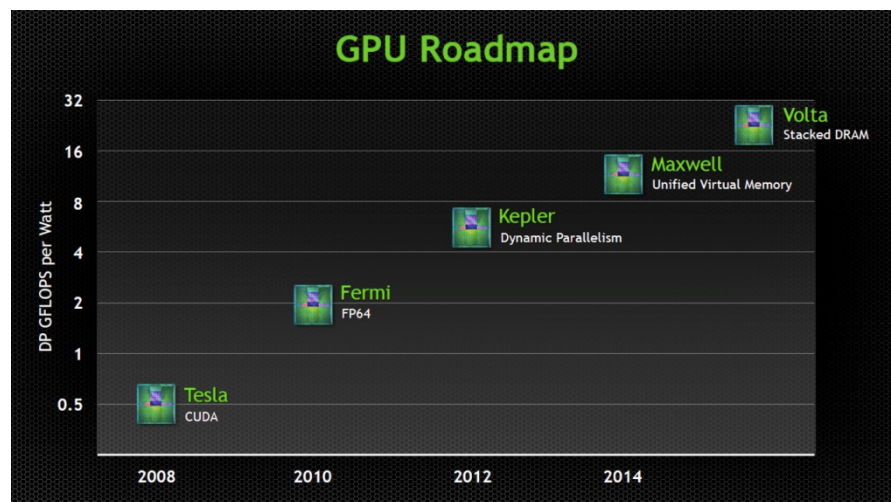
- Architetture many-core e il “**memory wall problem**” (oltre all’ “I/O wall problem”)
 - E’ tecnologicamente “piu’ semplice” scalare i TFlops (i.e. add cores...) che la banda verso la memoria (limitata dal numero e dalla frequenza di switching dei pins di I/O)
- GPGPU mostra efficienze computazionali limitate dalla banda passante verso la memoria locale con un rapporto $R = \text{GFlops}/\text{GB}$ non garantito di generazione in generazione
 - Fermi 1TFlops per 150 GB/s ($R=6.5$)
 - Kepler 4TFlops ma “solo” 250 GB/s ($R=16$)
- Questo implica che lo speed-up della tecnologia di rete-I/O e’ solo “parzialmente essenziale”
 - PCI Gen2/3 e’ sufficiente per sostenere l’I/O richiesto dalla media delle applicazioni



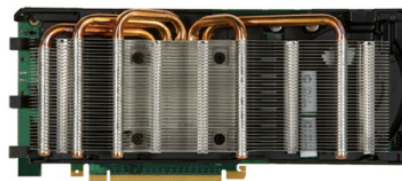
INTEL Westmere
+ many caches - few processing



Nvidia Fermi GPU
many computing units!!!

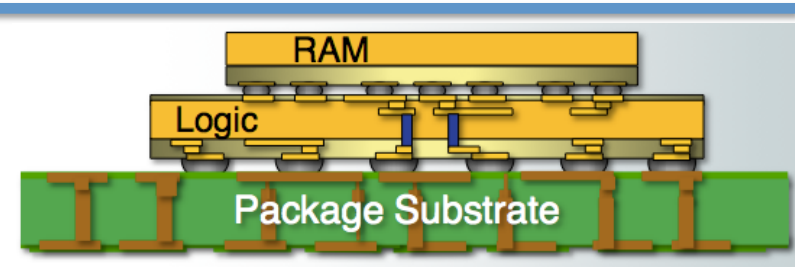


Nvidia Fermi (Tesla 20xx)
~500 core, **1 TF SP**, **0.5 TF DP**
6 GB external memory (**150 GB/s**)

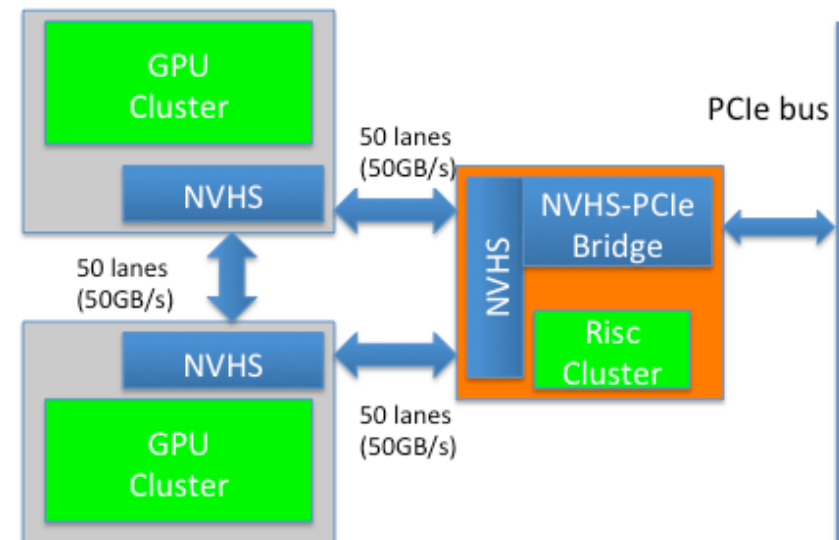
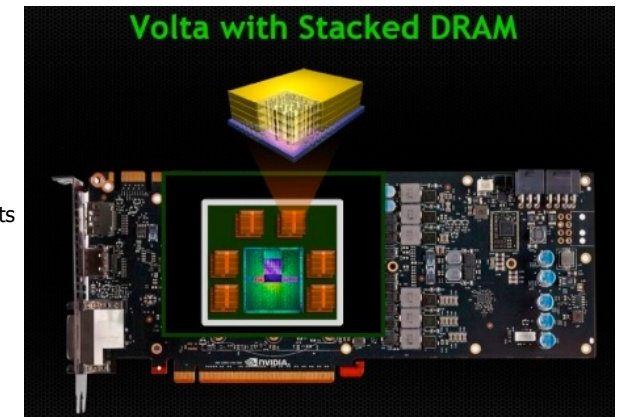


Nvidia Kepler (K20..)
~2500 CUDA core, **>3x FP** (4 TF(SP)/1.3TF (DP))
6 GB external memory (**250 GB/s**)

- Nel futuro (2016 ->...) vendors utilizzeranno nuove strategie architetturali e tecnologiche per ridurre l'impatto del "memory wall problem" e migliorare la bandwidth con la memoria locale
 - Integrazione di multiple RISC (ARM) cores on chip
 - "3D Integration" per realizzare "Stacked memory on chip"
 - In principio $\sim N^2$ pins (perimetro vs area) a frequenza molto piu' alta (no wiring...)
 - Nvidia **VOLTA** nel 2016, 4x in performance ma R=16
- La tendenza e' lo sviluppo di un canale proprietario ad altissime performances, on-chip e bridging verso standard di rete e/o di I/O.
 - Obiettivo primario: sistemi multi-gpu fortemente integrati
 - Rumours su NVIDIA GPU del 2016-17
 - **ORION** architecture (RIGEL+GM210s)
 - Cluster di N (24?) **ARM9(15) on chip**
 - N cuda Cores
 - **NVHS**: canale di I/O dedicato
 - >100 linee seriali indipendenti e clusterizzabili a gruppi
 - Inizialmente 10Gb/s per linea -> 100 GB/s
 - Scaling con speed-up della velocita' del link



Nvidia VOLTA
 Est. 4x K20x performance
 1 TB/s memory bandwidth
 New architectural improvements



APEnet+ roadmap nel 2013-2014

- APEnet+ Versione 5 (V5) caratterizzata da FPGA di nuova generazione 28nm (StratixV)
 - -> PCI Gen2 -> Gen3 & Esplorazione Gen3 x16
 - Bandwidth x2 verso host e GPU
 - -> Integrazione APEnet+ V5 con GPU di nuova generazione (Kepler e oltre)
 - Upgrade delle performance dei links del toro (8,5 Gb/s -> 14 Gb/s)
 - -> Interfaccia con i canali proprietari delle nuove/future GPU (?)
 - Nuova interfaccia PCI vs Backend APEnet basata su AMBA-AXI
 - -> Possibile integrazione APEnet – ARM
- Esplorazione di specializzazione di sistema a partire dai requisiti delle applicazioni stesse.
 - Routing function customizzabile
 - “Processing in network”: acceleratori hardware integrati nella FPGA e operanti sul flusso dati
 -

Thank you!

Any question or comment?