# News on
# Network Infrastructure and Fault Tolerance

U. Fuchs (PH-AID-DA)

13 March 2013

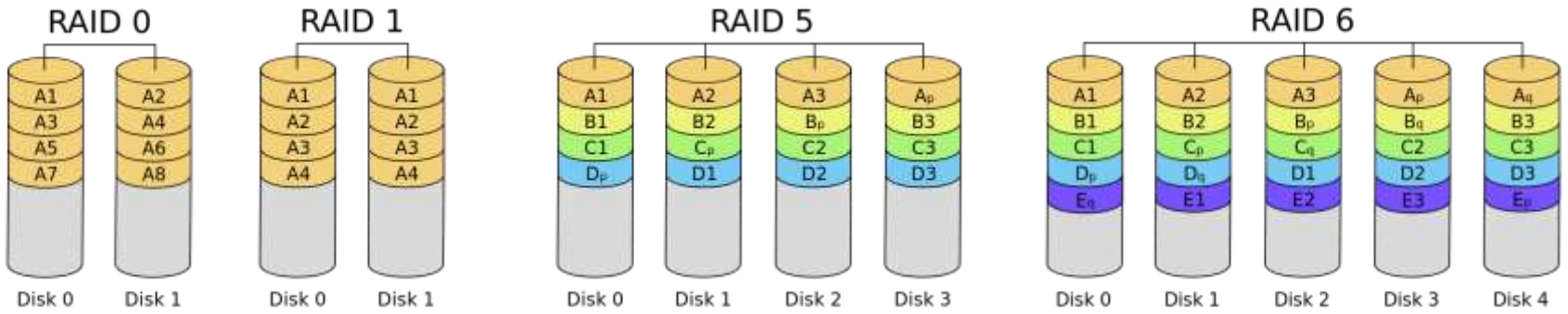
# Contents


- Fault-tolerant disk systems
  - RAIDs, and why they should be re-considered
  - DDP – Dynamic Disk Pools
- Data Center Networks,
  From Lossy to Loss-Less
  - Present situation
  - The Fiber Channel roadmap
  - Converged networks, DCB
  - FCoE

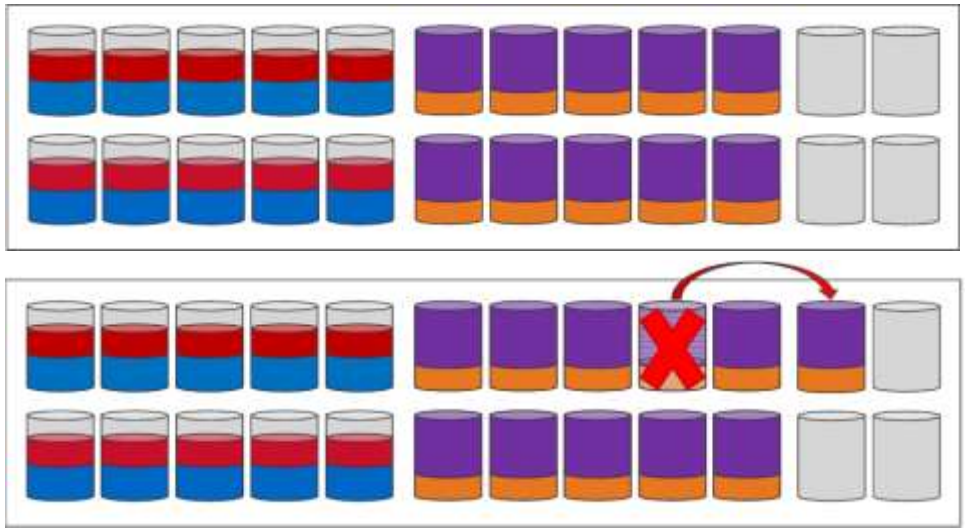
# RAID - DDP

- RAID (Redundant Array of Inexpensive Disks)
  - Combine multiple disk drives into a logical unit to increase the level of redundancy or performance, depending on the RAID level.

| RAID level | Description | Fault tolerance | Array failure rate | Read performance | Write performance |
|---|---|---|---|---|---|
| RAID – 0 | Block-level striping without parity or mirroring | 0 (none) | $1-(1-r)^n$ | nX | nX |
| RAID – 1 | Mirroring without parity or striping | n-1 | $r^n$ | nX | 1X |
| RAID – 5 | Block-level striping with distributed parity | 1 | $\frac{1}{2}n(n-1)r^2$ | (n−1)X | (n−1)X |
| RAID – 6 | Block-level striping with double distributed parity | 2 | $\frac{1}{6}n(n-1)(n-2)r^3$ | (n−2)X | (n−2)X |

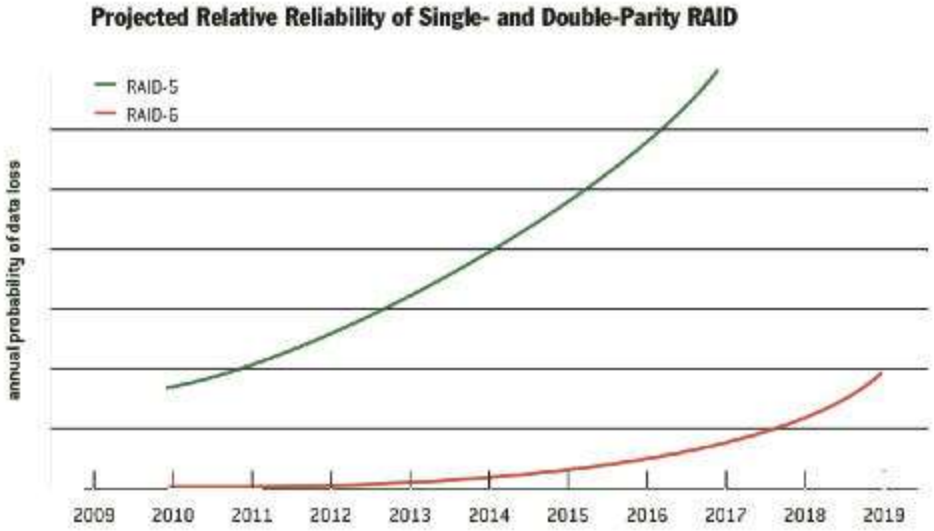


- RAID Failures
  - Single disk-failure
    - Content of failed disk is re-constructed from Parity stripe (drive rebuild) and a placed on a spare disk



  - Double disk-failure
    - In RAID-5: data lost
    - In RAID-6: second Parity stripe can save the situation
      - In case of a single bit error during read, the rebuild will fail. Data lost.

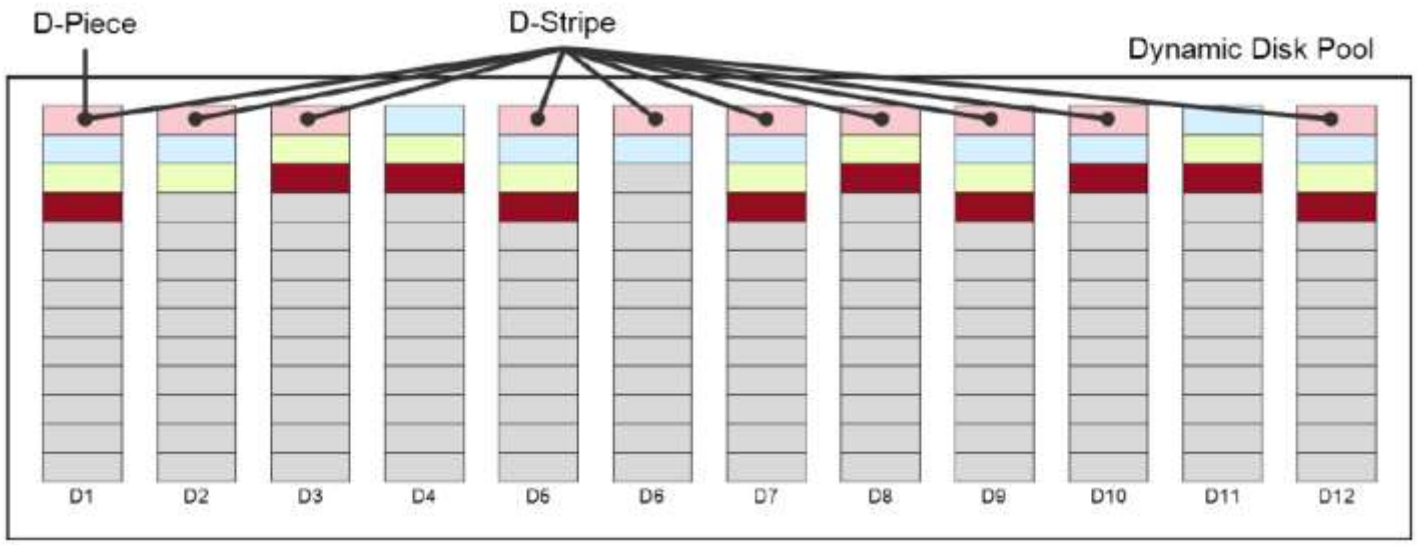
- RAID Failures
  - RAID drive rebuilds take a long time
    - 8x 350GB drives: ~ 1-2 hours
    - 8x 1TB drives: ~ 1 day
    - 8x 3TB drives: ~3 days
    - Soon: 8x 20TB drives: ... ~ 1 month, 16x20TB: ~ 2 months, …
  - Further drive failures during rebuild are probable (disks running at 100%)
    - Rebuild failure (for RAID5) or double-rebuild (up to time x 4) for RAID-6




Courtesy ACM Queue

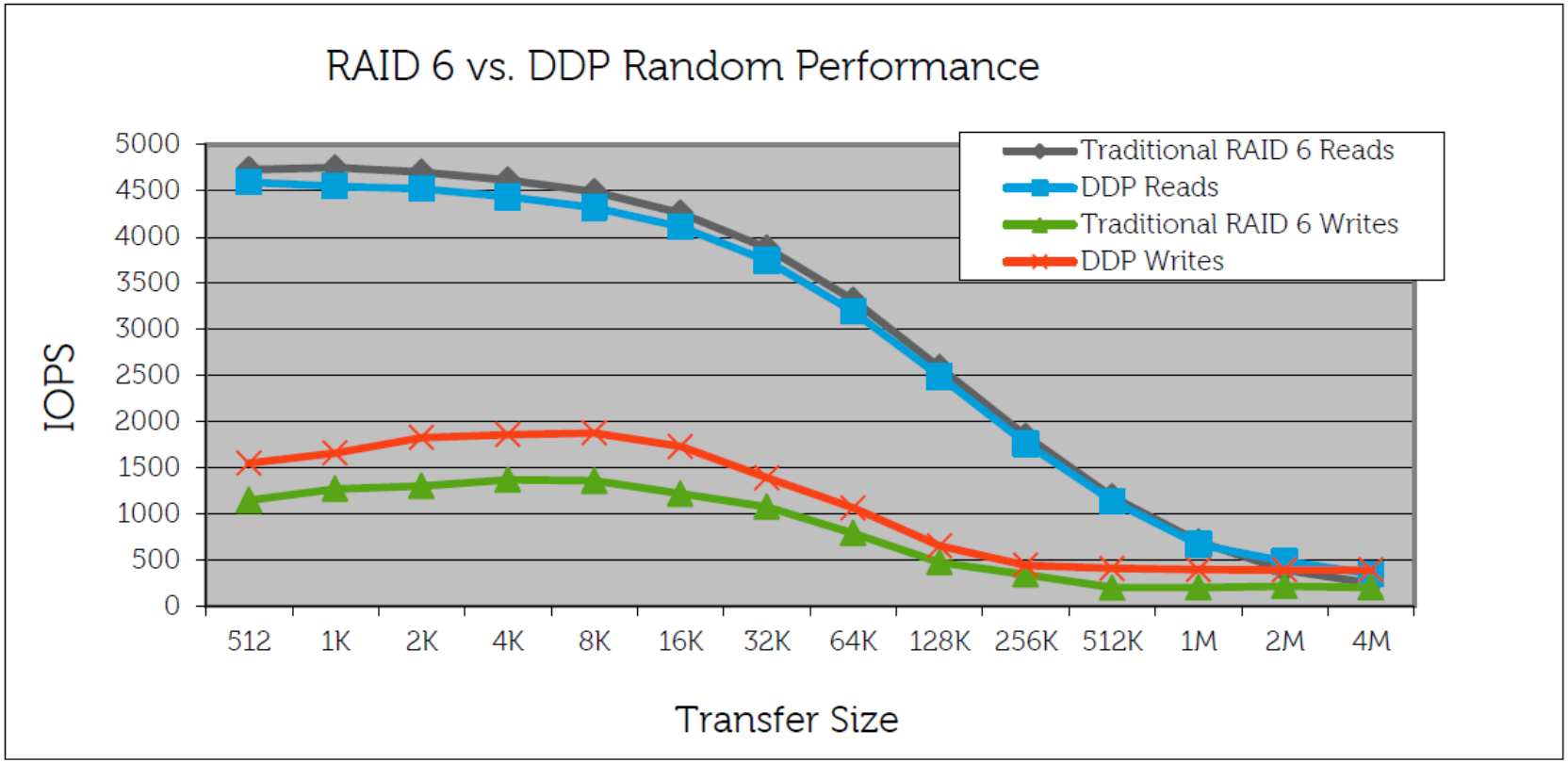
  - Something has to change.

- DDP – Dynamic Disk Pools
  - Each data-stripe is written to some disks (not all) as data-piece
  - Two parity pieces, no spare drives

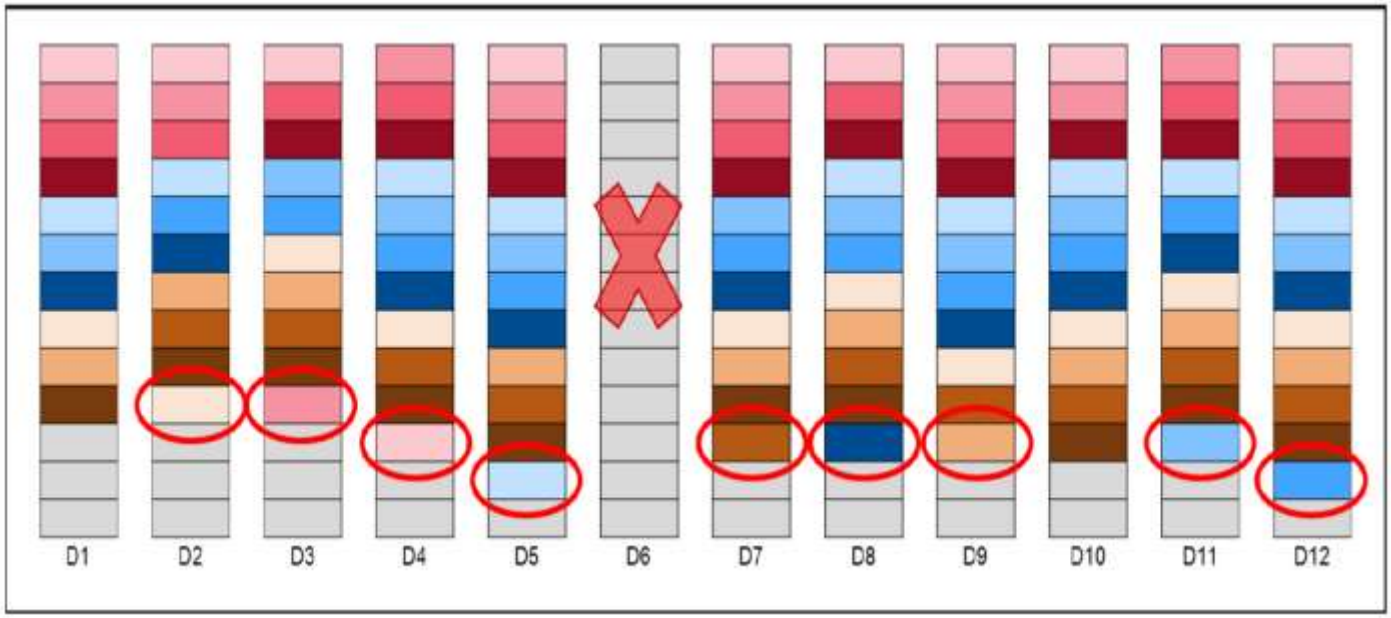



  - Immediate Gain:
    - Space (no spare drives)
    - Performance (not all drives are used, less contention)
    - Due to non-uniform usage: easily re-sizable on-line, no rebuild necessary.

- DDP – Dynamic Disk Pools
  - Performance comparison with RAID6:

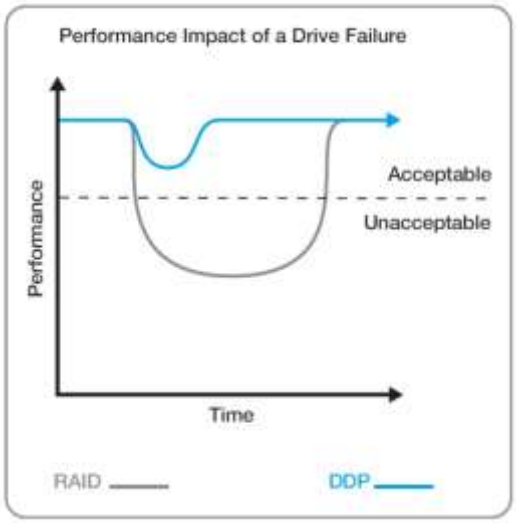


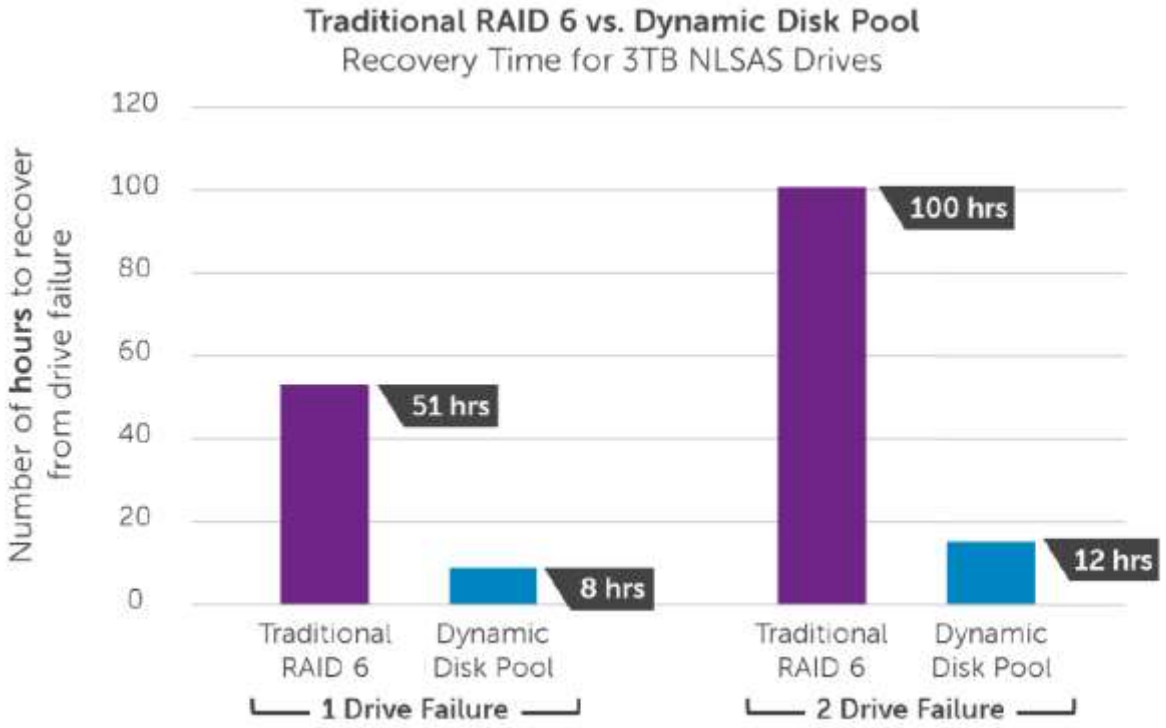
- DDP – Dynamic Disk Pools
  - In case of failure, the missing D-Pieces are recalculated and appended to working disks (avoiding two D-Pieces of the same stripe on the same disk).




  - Advantage:
    - Rebuild is fast (many disk read, many disk write) [RAID: many-to-one]
    - Non-affected volumes stay available [RAID: whole volume heavily impacted]
    - General performance drop is much lighter than in RAID

- DDP – Dynamic Disk Pools
  - Rebuilding …



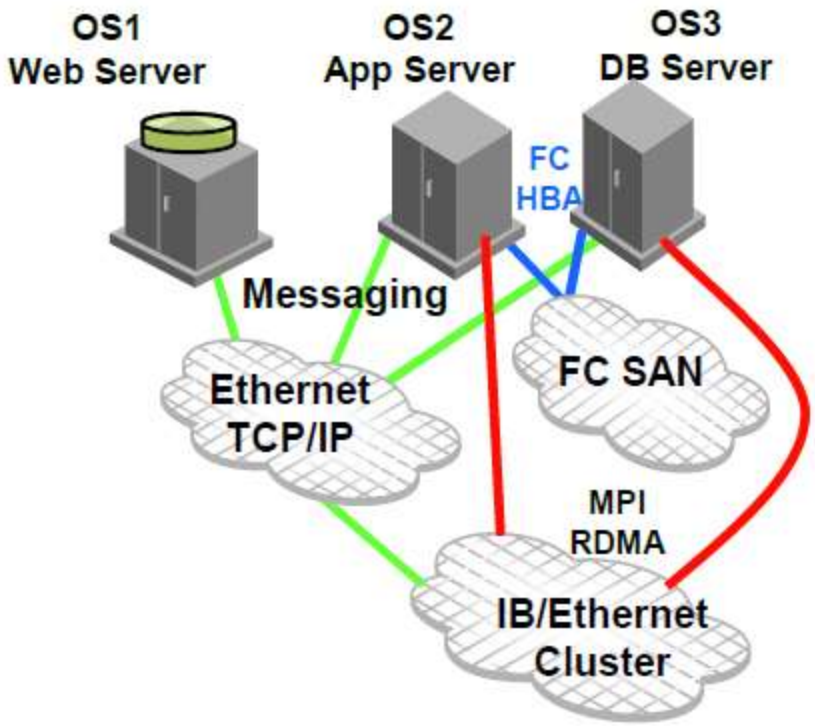
- RAID ??

# Trends in Datacenter networks,

## From Lossy to Loss-less

### From Plug-fests to a Converged Infrastructure

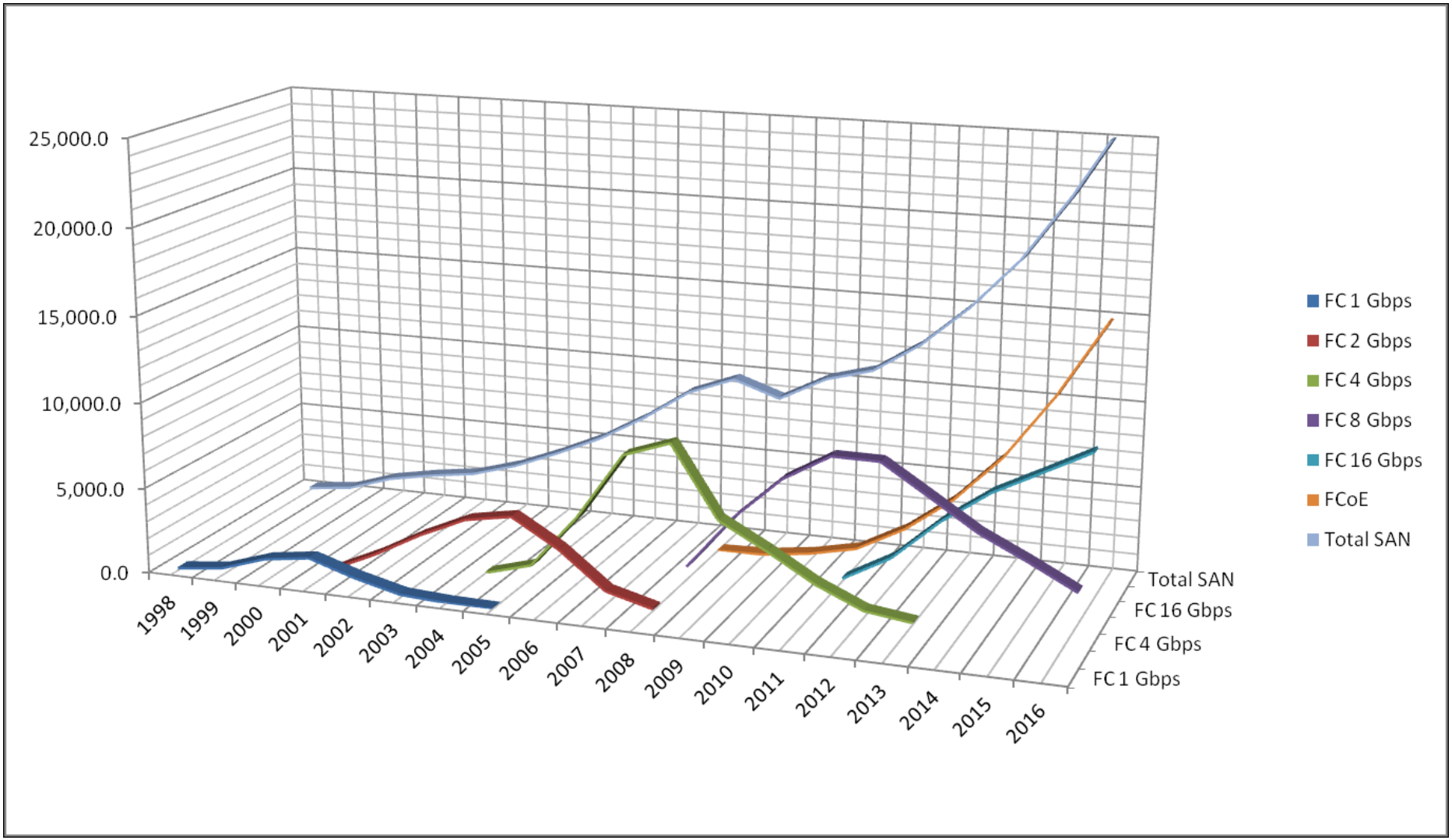
- Datacenter networks
  - Situation today
    - Servers are connected to several networks
    - … have several network interfaces (+software stacks)
    - … provide interesting cabling challenges
  - Money spent on
    - Adapters
    - Switch ports
    - Power
    - Cooling
    - .. Operations, Supervision
  - Most prominent players:
    - Fiber Channel for storage
    - Ethernet TCP/IP for network access



- FiberChannel Roadmap
  - FC clearly dominates the SAN market
  - Well understood, convenient to use and implement
  - Continuous speed and Bandwidth/$ improvements
  - Aggressively pursuing Energy Efficiency (best Efficiency/Watt rating)

| Product Naming | Throughput (MBps) | Line Rate (GBaud) | T11 Spec Technically Completed (Year)‡ | Market Availability (Year)‡ |
|---|---|---|---|---|
| 1GFC | 200 | 1.0625 | 1996 | 1997 |
| 2GFC | 400 | 2.125 | 2000 | 2001 |
| 4GFC | 800 | 4.25 | 2003 | 2005 |
| 8GFC | 1600 | 8.5 | 2006 | 2008 |
| 16GFC | 3200 | 14.025 | 2009 | 2011 |
| 32GFC | 6400 | 28.05 | 2012 | 2014 |
| 64GFC | 12800 | TBD | 2015 | Market Demand |
| 128GFC | 25600 | TBD | 2018 | Market Demand |
| 256GFC | 51200 | TBD | 2021 | Market Demand |
| 512GFC | 102400 | TBD | 2024 | Market Demand |

- FiberChannel Roadmap



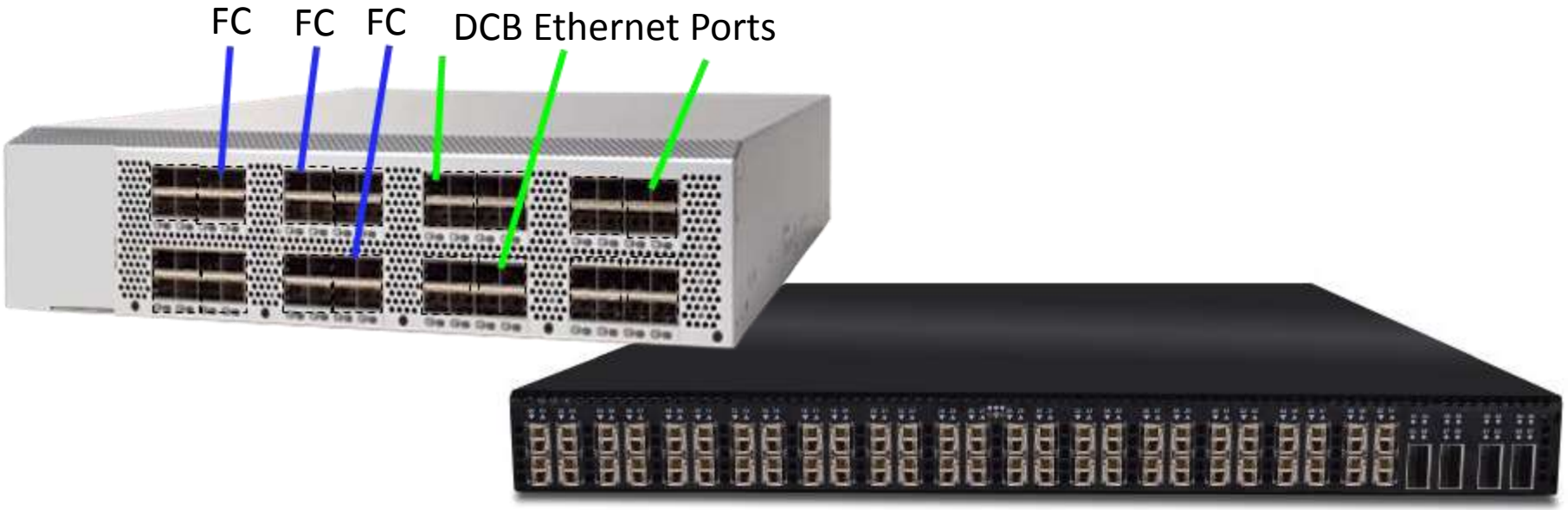
- FiberChannel Roadmap
  - FCoE (Fiber Channel over Ethernet) standards available since 2008
  - Encapsulation of FC in Ethernet; another upper-layer protocol
  - Same cabling (SFP+) for 8G FC, 16G FC and 10G FCoE

| Product Naming | Throughput (MBps) | Equivalent Line Rate (GBaud)† | Spec Technically Completed (Year)‡ | Market Availability (Year)‡ |
|---|---|---|---|---|
| 10GFCoE | 2400 | 10.3125 | 2008 | 2009 |
| 40GFCoE | 9600 | 41.225 | 2010* | Market Demand |
| 100GFCoE | 24000 | 103.125 | 2010* | Market Demand |

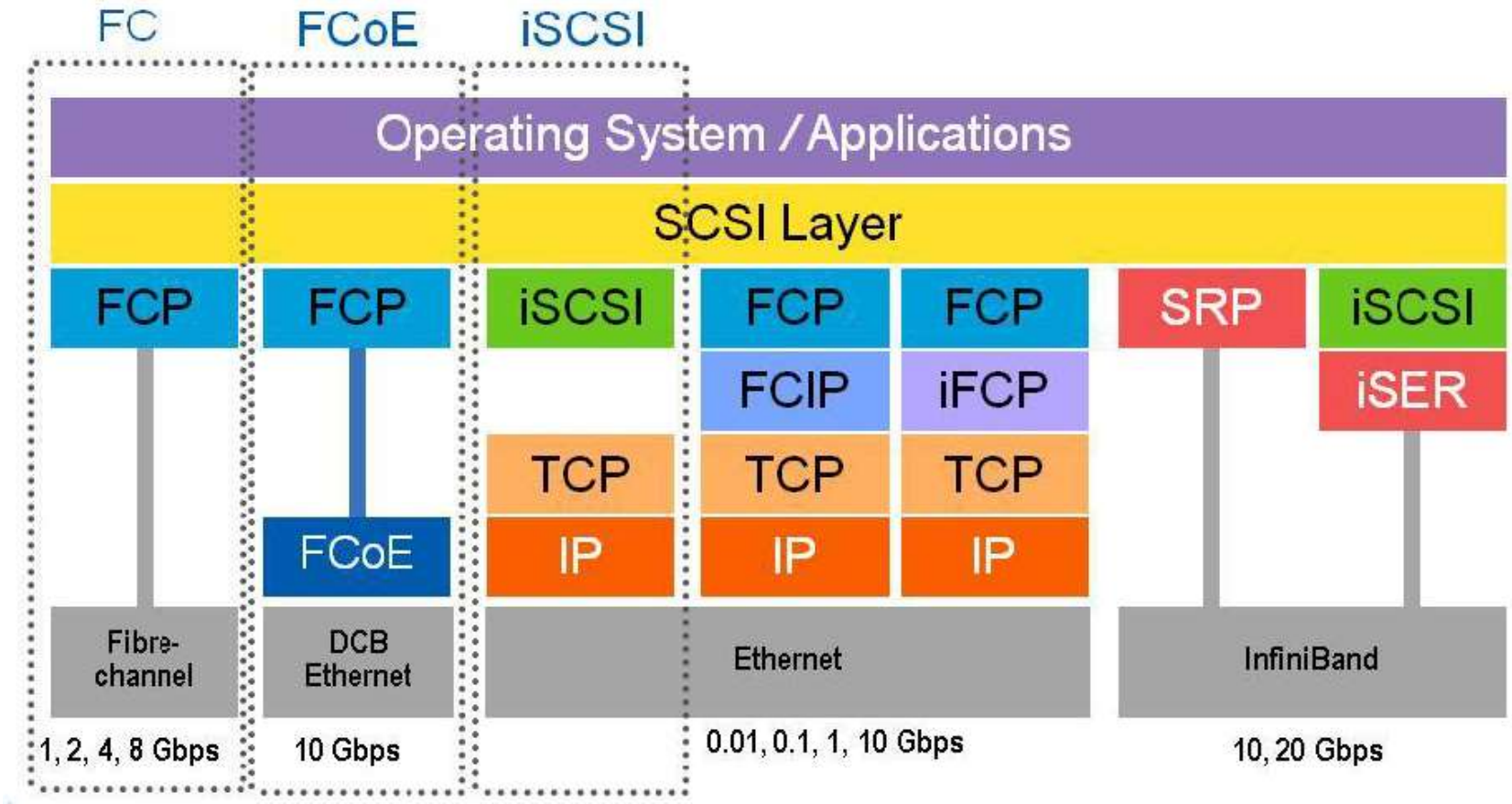
- FCoE challenge
  - FCoE is the encapsulation of FC in Ethernet


  - The Challenge:
    - FC (by design) is a loss-less protocol
    - Ethernet (by design) is a lossy protocol
  - The Solution:
    - We need loss-less Ethernet with extensions: **DCB (Data Center Bridging)**
      - Advances in Ethernet recently defined in IEEE 802.1, specifically:
        - Priority-based Flow Control (PFC) 802.1Qbb
        - Enhanced Transmission Selection (ETS) 802.1Qaz
        - DCB (capability) eXchange (DCBX) Protocol 802.1Qaz
        - CN -- Congestion Notification (802.1Qau)
        - Possible future Multi-pathing (IETF– TRILL, IEEE 802.1aq-SPB, et.al.)
    - FCoE requires these Ethernet extensions to be implemented, Lossless switches and fabrics (e.g., supporting IEEE 802.3 PAUSE), Jumbo frame support is strongly encouraged

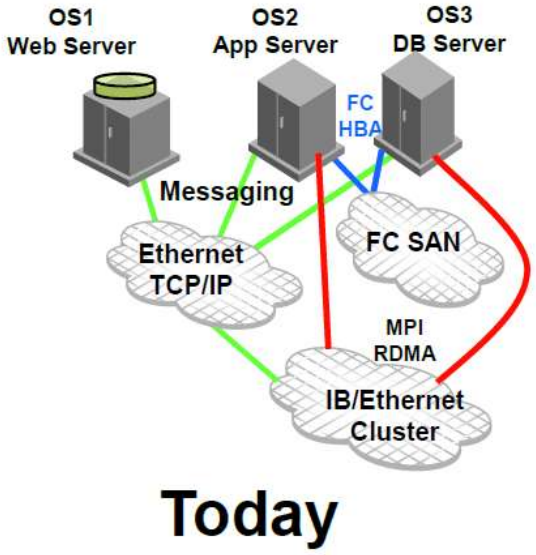
- DCB
  - FCoE Fabrics must be built with DCB Switches that:
    - Are called Fiber Channel Forwarder (FCF)
    - Are part of a lossless Ethernet Fabric and have DCB Lossless Ethernet ports
    - Support Ethernet and IP standards for switching, pathing and routing
    - Support FC standards for switching, pathing and routing
    - Adapt between FCoE, FC and Ethernet

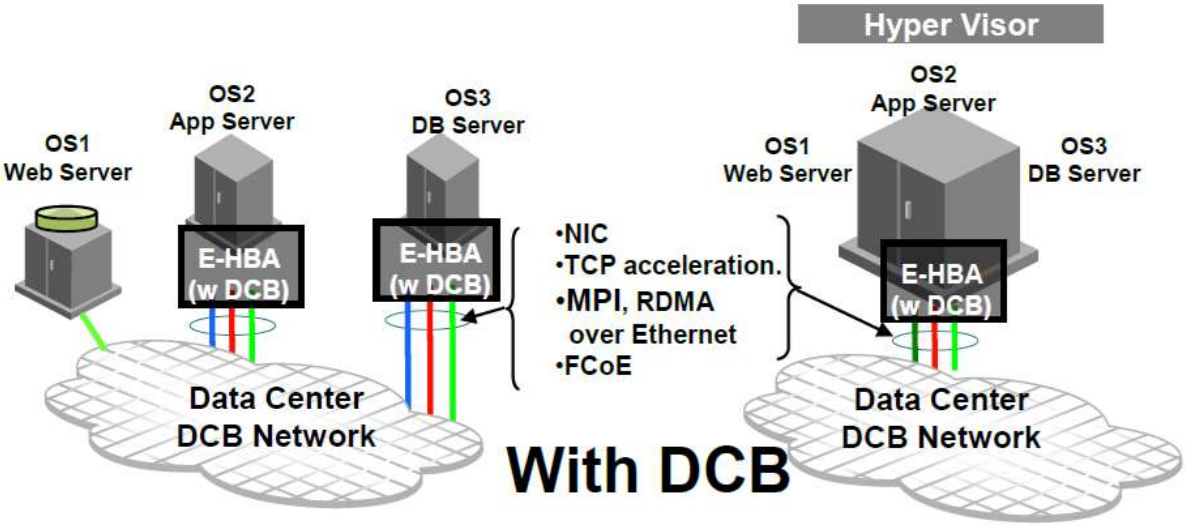


- DCB stack


FC
FCoE
iSCSI
Operating System / Applications
SCSI Layer
FCP
FCP
iSCSI
FCP
FCP
SRP
iSCSI
FCIP
iFCP
iSER
TCP
TCP
TCP
FCoE
IP
IP
IP
Fibre-channel
DCB Ethernet
Ethernet
InfiniBand
1, 2, 4, 8 Gbps
10 Gbps
0.01, 0.1, 1, 10 Gbps
10, 20 Gbps

- Datacenter networks
  - Dramatic Interface reduction in adapters, switch ports, cabling, power and cooling
  - 4-6 cables can be reduced to 2 Interfaces/cables per server
  - Seamless connection to the installed base of existing SANs and LANs
  - Effective sharing of high bandwidth links



Thank you.