Universitat Oberta de Catalunya

Simular entorns HPC amb aplicacions MPI per avaluar el rendiment fent ús de diferents tecnologies i millores

Treball Final de Màster

Autor: José Antonio Martín Pérez
Consultor: Alberto García Villoria

Gener 2016

Índex

- Introducció
- Motivació
- Objectius
- Conceptes teòrics
  - Tecnologies d'interconnexió
  - Topologies
- Modelatge i simulació
  - SimGrid
  - MPI
- Sistema base
- Millores proposades
- Avaluació de les millores
- Conclusions
- Treball futur

Introducció

Màster en Enginyeria Informàtica

Treball Final de Màster

Àrea
Simulació i Modelatge

Conceptes clau
HPC , MPI , SimGrid
Fat Tree , Torus , Ethernet , InfiniBand

Motivació

- Necessitats de càlcul intensiu:
  Recerca, Business Intelligence, Big Data, Cloud
- Millores tecnològiques constants
- Necessitat d'avaluar aquestes
- Dificultat d'anàlisi
  sense usar simulació

Objectius

- Modelar sistema HPC amb *Simgrid*
- Definir característiques del sistema:
  - Tecnologies d'interconnexió: *Ethernet* e *InfiniBand*
  - Topologia: *Fat Tree* , *2D Torus* i *3D Torus*
- Simular sistema base amb programari MPI
- Aplicar diferents millores sobre sistema base
- Avaluar el rendiment aconseguit
- Extreure conclusions

Conceptes teòrics

- Tecnologies d'interconnexió:
  - Ethernet: Gigabit Ethernet i 10 Gigabit Ethernet
  - InfiniBand: SDR , DDR , QDR , FDR i EDR
- Topologies clàssiques en entorns HPC:
  - Fat Tree
  - 2D Torus
  - 3D Torus
- Programari per simular entorns HPC: SimGrid
- Programació paral·lela: MPI

Tecnologies d'interconnexió

- Gigabit Ethernet
  - Molt utilitzat actualment en entorns TIC
  - Preu molt econòmic
  - Bona relació preu / rendiment
  - Present a entorns d'alt rendiment (HPC)
  - Velocitat de transmissió: 1.000 Mbps
  - Latència: 4,25 µs
- 10 Gigabit Ethernet
  - Evolució del estàndard anterior
  - Preu cada cop més assequible
  - Millora important del rendiment (10x)
  - Augment velocitat transmissió / Reducció de latència
  - Velocitat de transmissió: 10.000 Mbps
  - Latència: 2,25 µs / 0,3 µs

Tecnologies d'interconnexió (2)

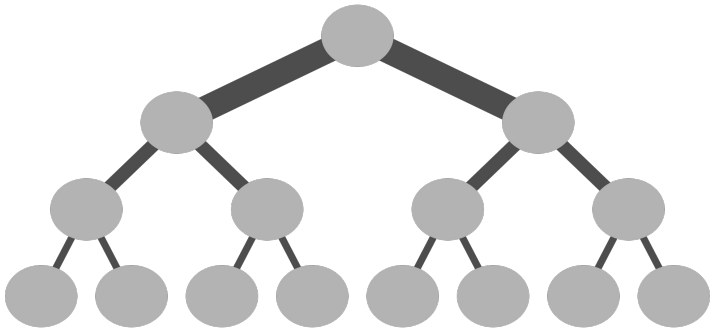
- 40 / 100 Gigabit Ethernet
  - Última versió de l'estàndard Ethernet
  - Apareix l'any 2010
  - Augment de velocitat important (4x i 10x)
  - Mateixa latència: 0,3 µs
  - Cost encara elevat
  - Es defineixen dues velocitat de transmissió:
    - 40 Gbps per als nodes
    - 100 Gbps per connexions troncals

Tecnologies d'interconnexió (3)

- InfiniBand
  - Apareix l'any 1999
  - Desenvolupada específicament per entorns HPC
  - Objectiu: alta velocitat transmissió / baixa latència
  - Cost elevat i baixa demanda (entorns massa específics)
  - Fins l'actualitat, s'han definit cinc revisions:
    - SDR Velocitat: 10 Gbps / Latència: 5 µs
    - DDR Velocitat: 20 Gbps / Latència: 2.5 µs
    - QDR Velocitat: 40 Gbps / Latència: 1.3 µs
    - FDR Velocitat: 56 Gbps / Latència: 0.7 µs
    - EDR Velocitat: 100 Gbps / Latència: 0.5 µs

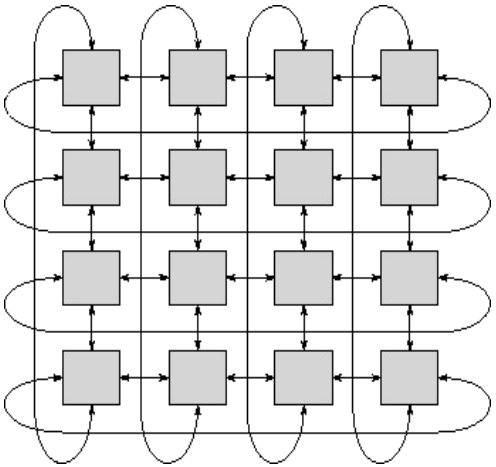
Topologies

- Fat Tree
  - Desenvolupada al MIT
  - Molt utilitzada al món de les comunicacions
  - Facilitat en el desplegament i l'escalatge
  - Necessitat d'administrar la xarxa: encaminadors (switches)
  - Basada en arbre amb diferents nivells
  - Rendiment basat en la relació entre nivells (1:1 – Òptim)



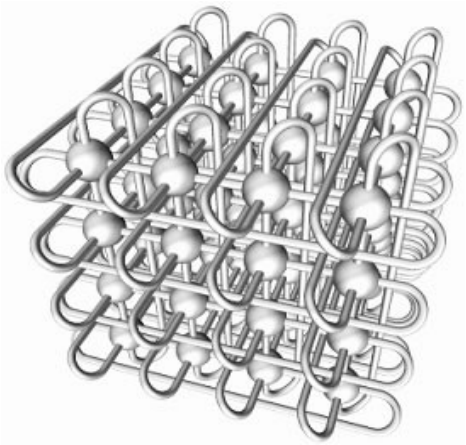
Topologies (2)

- 2D Torus
  - Específica en entorns d'alt rendiment (HPC)
  - Matriu de dues dimensions (N x N)
  - Nodes connectats directament, sense encaminadors
  - Difícil escalat: necessitat de mantindre l'equilibri de la matriu



Topologies (3)

- 3D Torus
  - Evolució de la topologia anterior
  - Molt utilitzada en entorns HPC
  - Cub de de tres dimensions (N x N x N)
  - Cada node connecta amb sis nodes veïns
  - Difícil d'escalar sense trencar l'equilibri del cub



Modelatge i Simulació

- SimGrid (http://simgrid.gforge.inria.fr/)
  - Desenvolupat a la Universitat de Califòrnia (San Diego)
  - Inici del projecte: 1999
  - Objectius: Simular entorns Grid, Cloud, HPC i P2P
  - Possibilitat de modelar sistemes HPC complexos
    - Ús d'un únic fitxer XML per modelar el sistema
    - Sintaxi senzilla e intuïtiva
  - Compatible amb programari MPI
    - Compilador propi: *smpicc*
    - Execució pròpia: *smpirun*

Modelatge i Simulació (2)

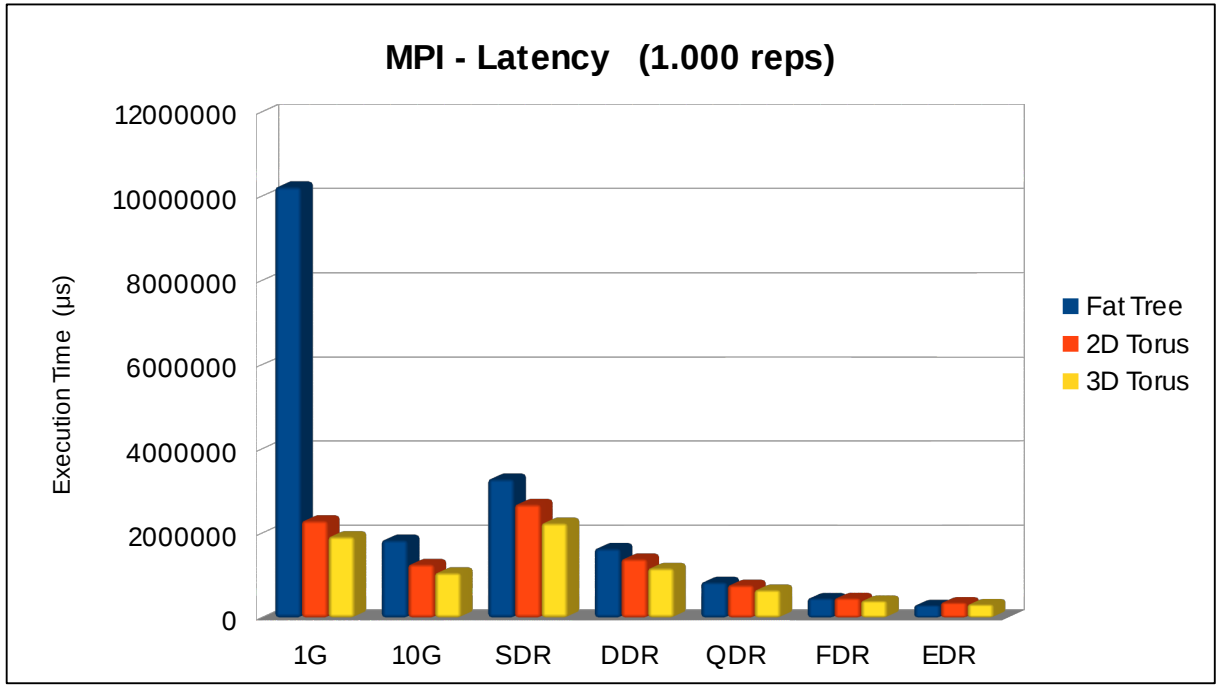
- MPI (*Message Passing Interface*)
  - Conjunt de llibreries per programació paral·lela
  - Compatible amb Fortran, C, C++ i Java
  - Molt utilitzat en sistemes HPC (memòria distribuïda)
  - S'ha fet ús de cinc aplicacions (llenguatge C i MPI):
    - Test de latència
    - Descomposició d'un vector gran (960M elements)
    - Multiplicació de dues matrius de mida gran
    - Generador de nombres primers
    - Càlcul equació d'ona

Sistema Base

- Nº de nodes: 64
- Tecnologies d'interconnexió
  - Ethernet: Gigabit Ethernet , 10 Gigabit Ethernet
  - InfiniBand: SDR , DDR , QDR , FDR , EDR
- Topologies
  - Fat Tree: 2 nivells , 4 Egde switch , 1 Core switch
  - 2D Torus: Malla de dues dimensions (8 x 8 elements)
  - 3D Torus: Cub de tres dimensions (4 x 4 x 4 elements)
- Processador als nodes
  - Intel Xeon E5-2687W
  - Rendiment de càlcul: 150 GFLOPS

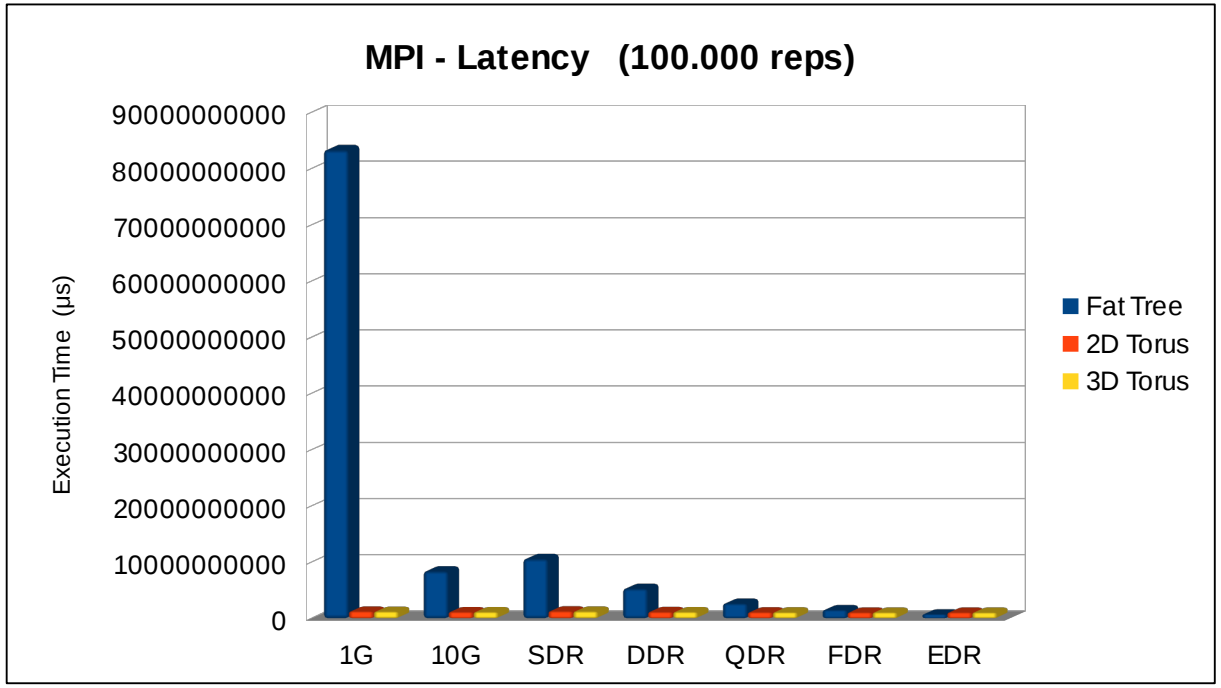
Avaluació del Sistema Base

- Test de latència (1.000 repeticions)



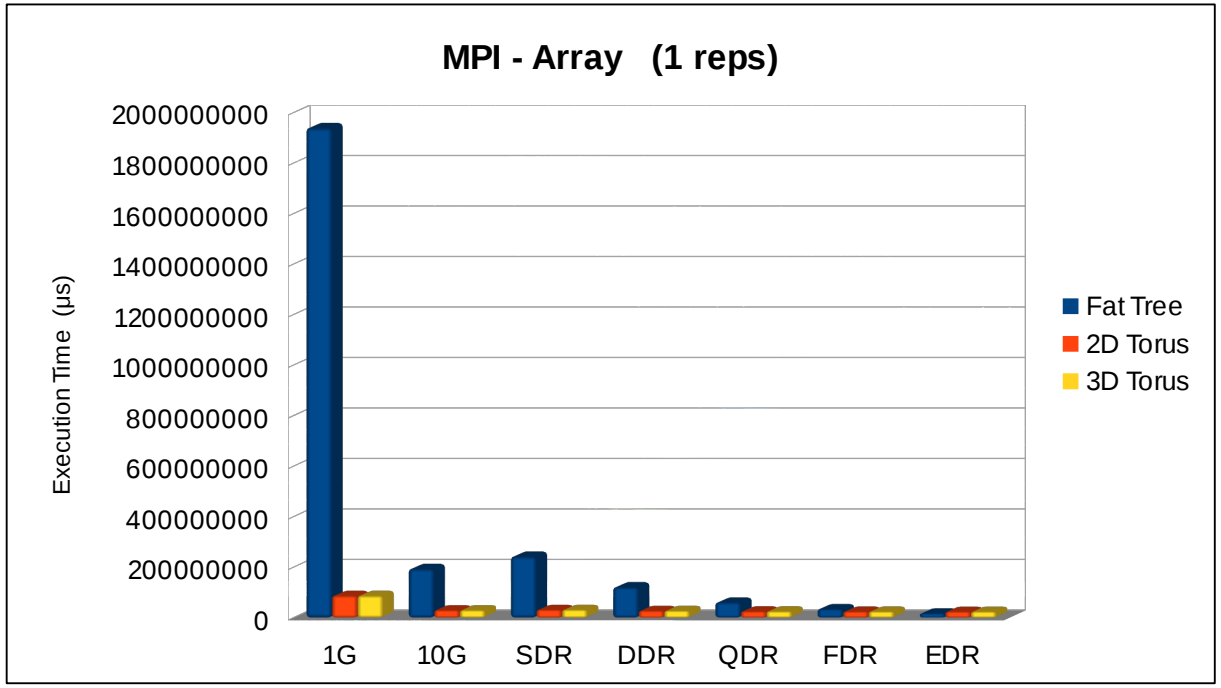
Avaluació del Sistema Base (2)

- Test de latència (100.000 repeticions)



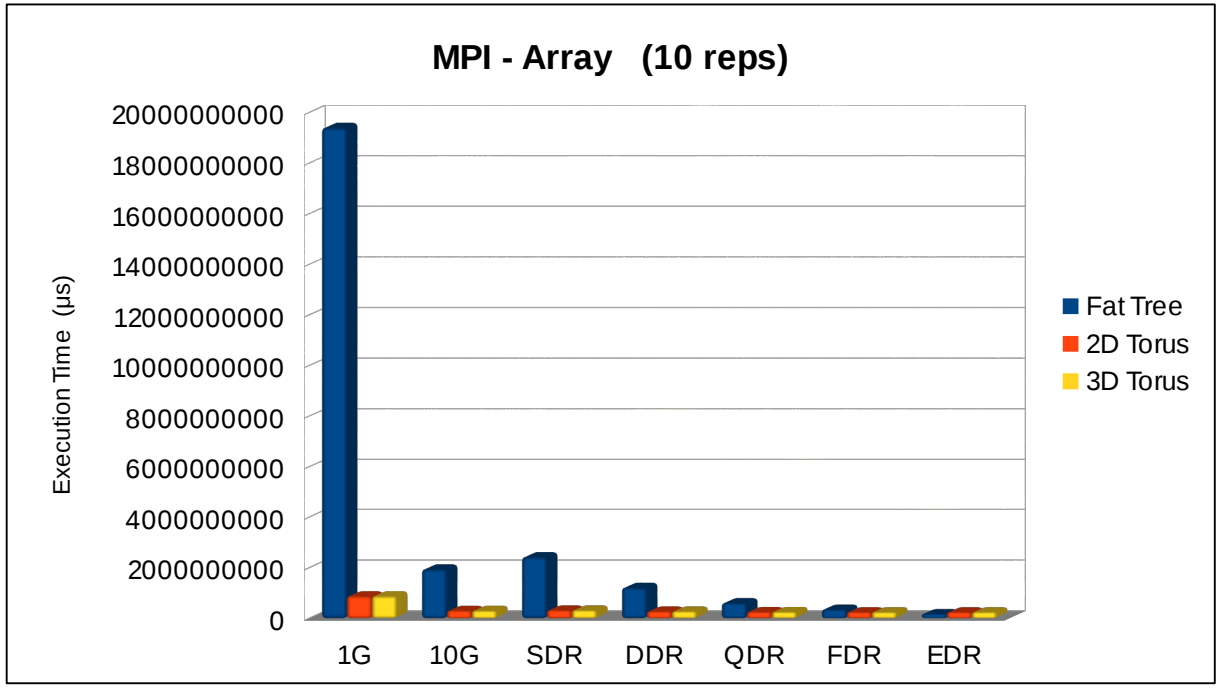
Avaluació del Sistema Base (3)

- Array Assignment (1 repetició)



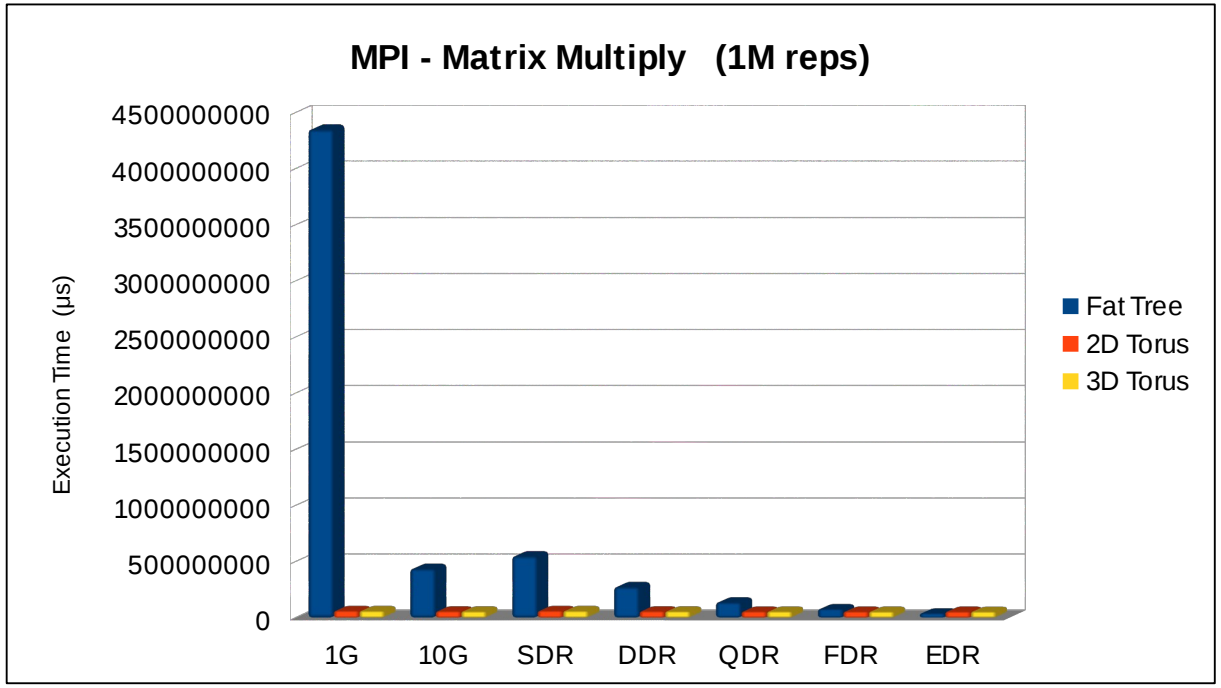
Avaluació del Sistema Base (4)

- Array Assignment (10 repeticions)



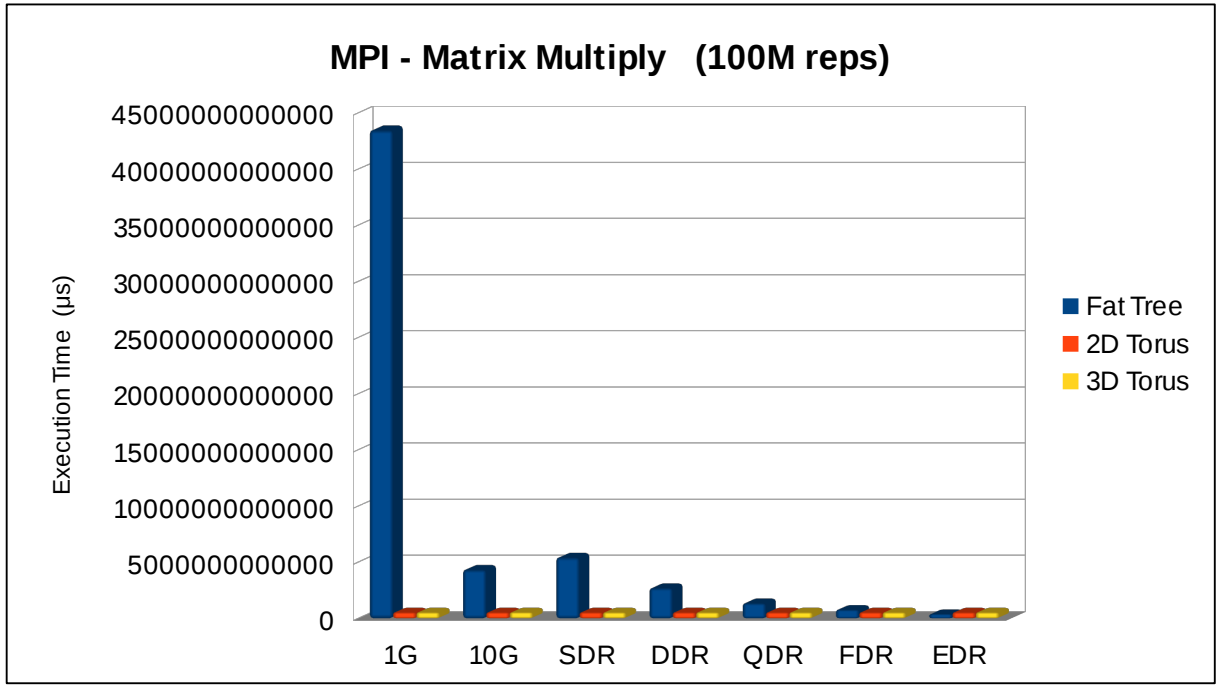
Avaluació del Sistema Base (5)

- Matrix Multiply (1.000.000 repeticions)



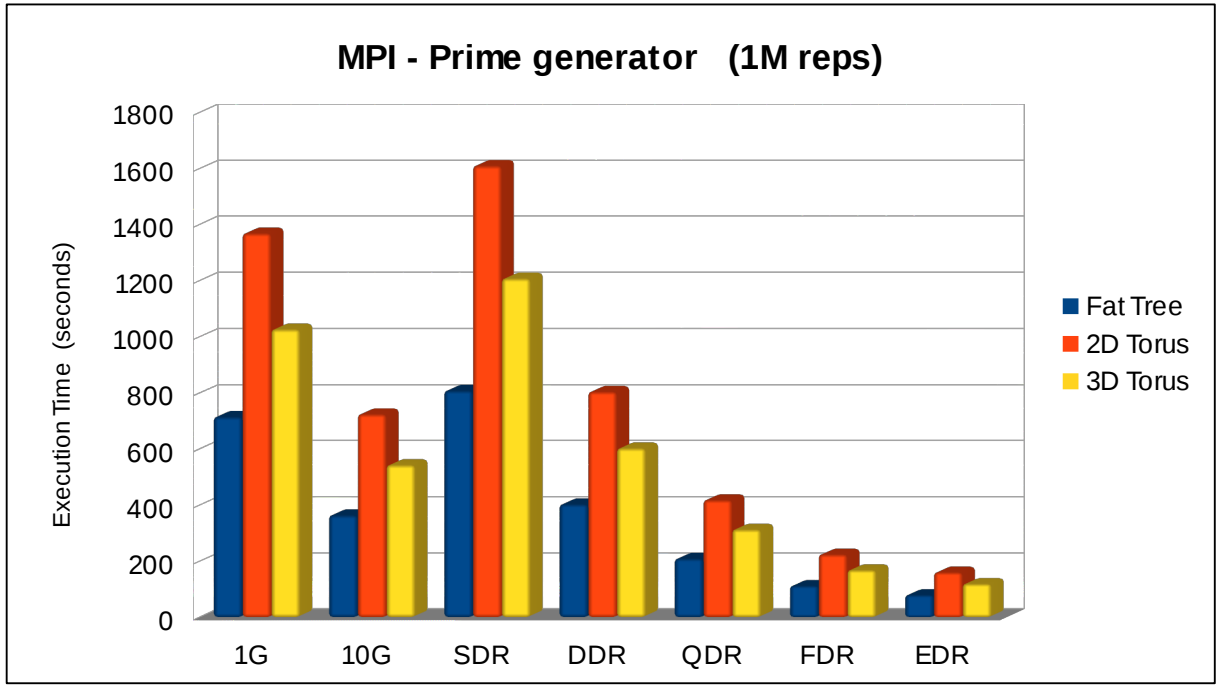
Avaluació del Sistema Base (6)

- Matrix Multiply (100.000.000 repeticions)



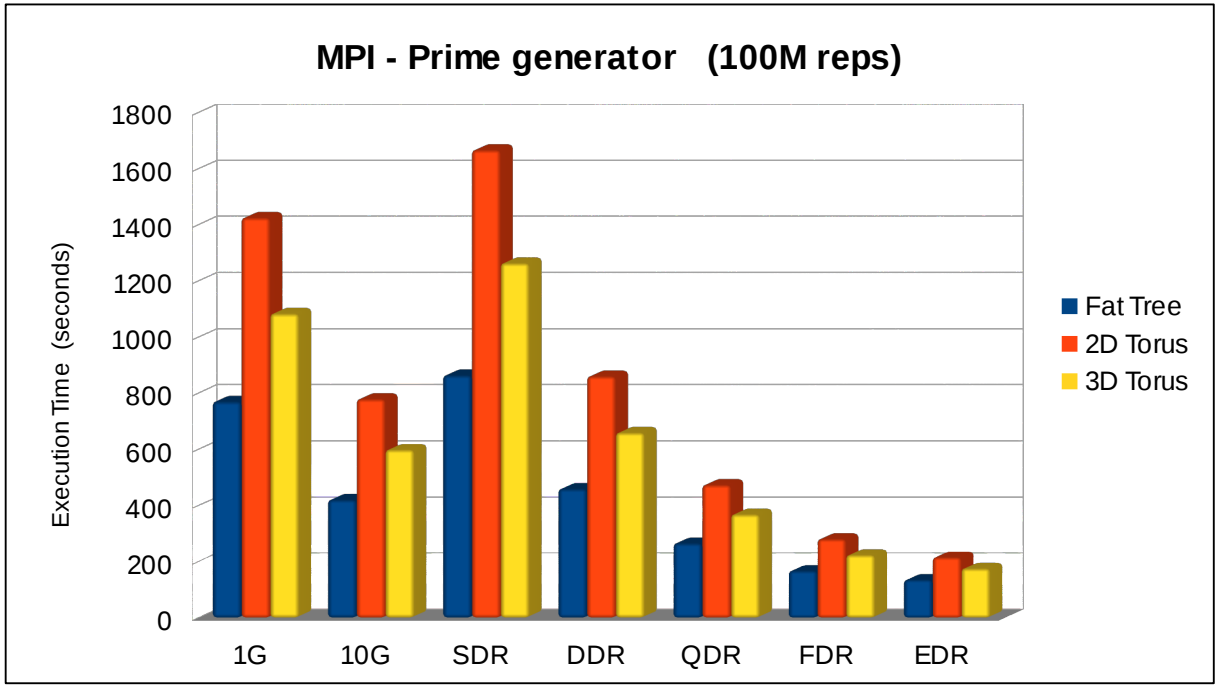
Avaluació del Sistema Base (7)

- Prime Generator (1.000.000 números generats)



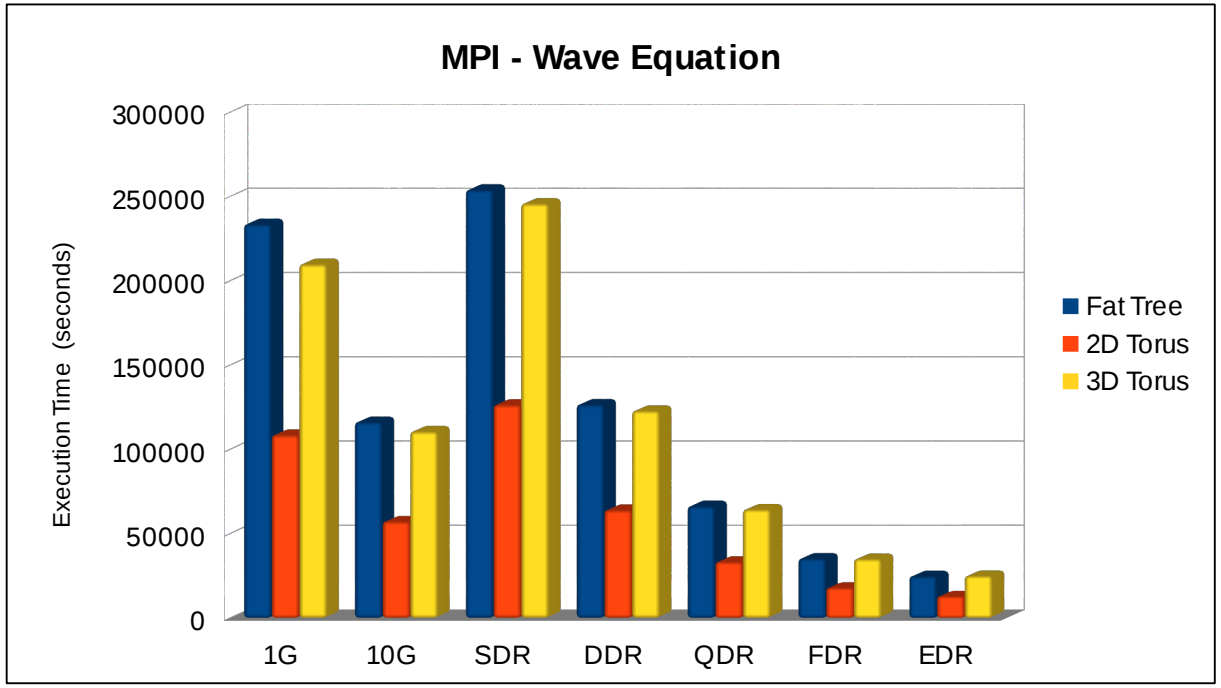
Avaluació del Sistema Base (8)

- Prime Generator (100.000.000 números generats)



Avaluació del Sistema Base (9)

- Càlcul equació d'ona (*Wave Equation*)

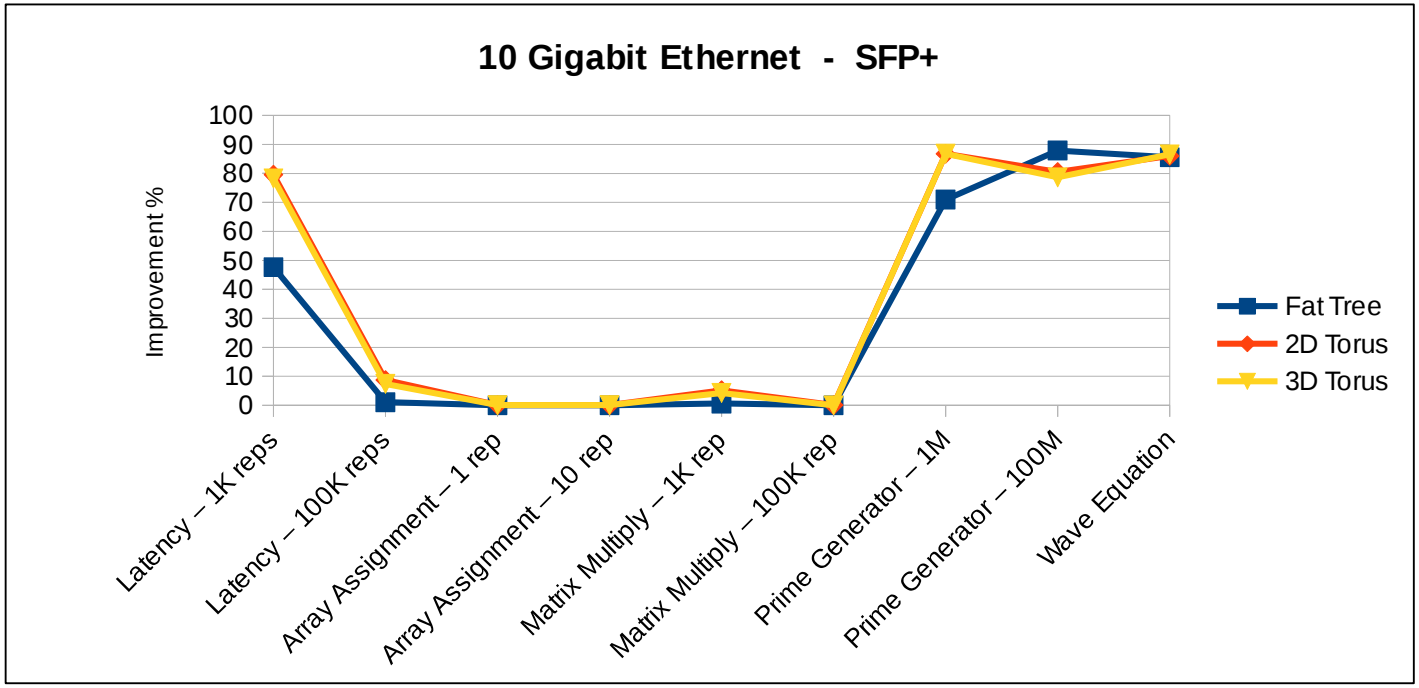


Millores proposades

- Tecnologies d'interconnexió
  - 10 Gigabit Ethernet amb SFP+
  - 40 / 100 Gigabit Ethernet
  - InfiniBand NDR
- Topologies
  - Fat Tree (1:1) amb InfiniBand
  - 2D Torus millorat
  - 3D Torus millorat

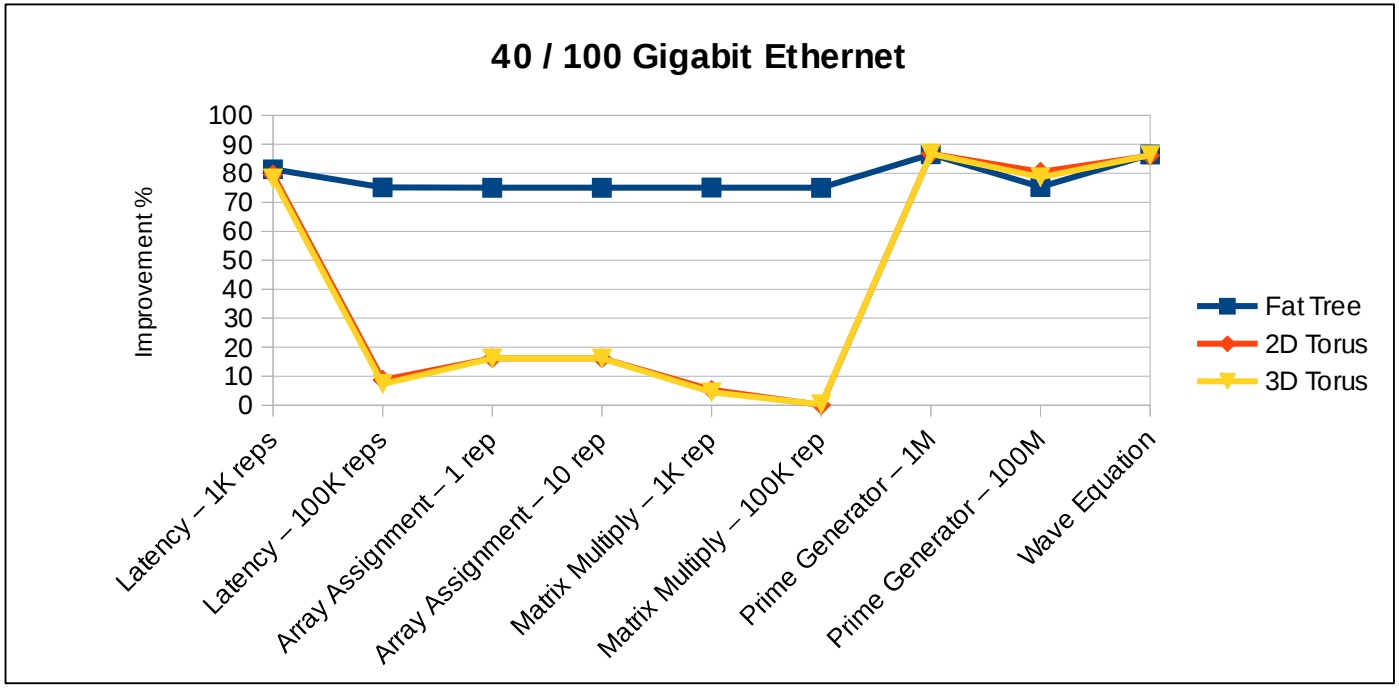
Avaluació de les millores

- 10 Gigabit Ethernet amb SFP+
  - Mateixa velocitat (10 Gbps) / Menor latència (0,3 µs)



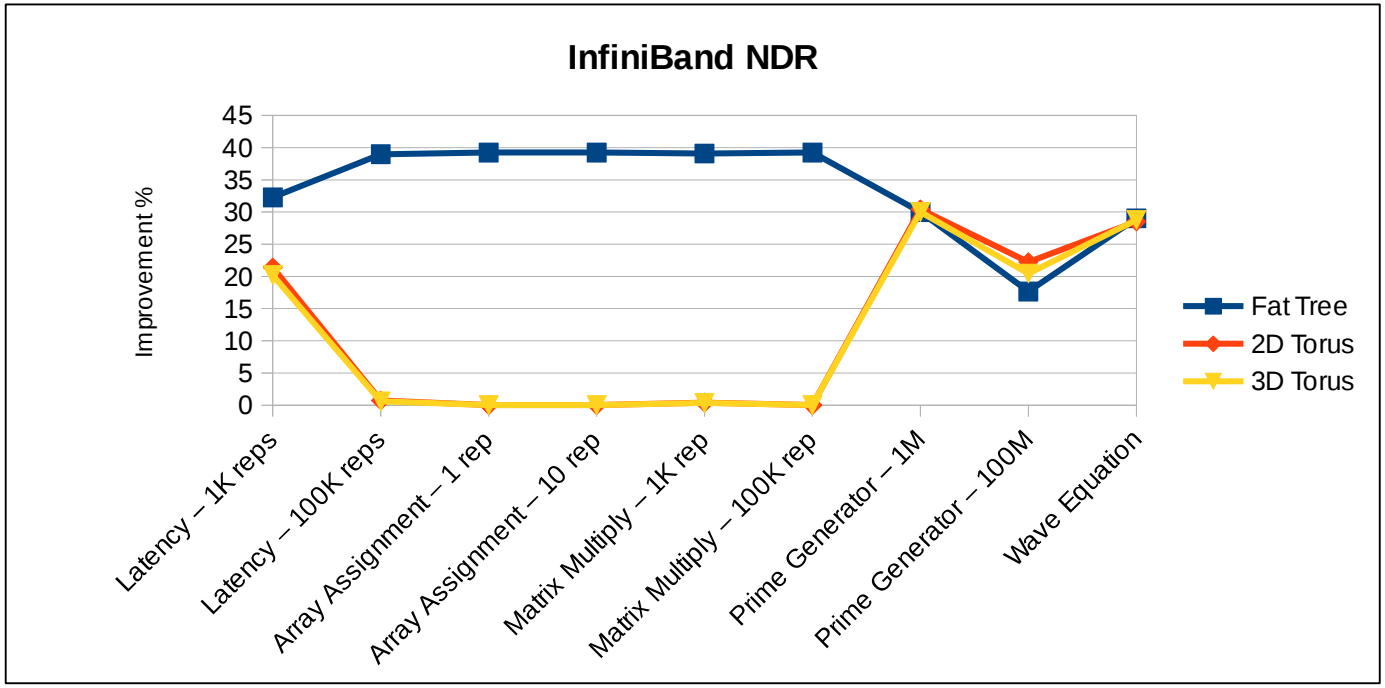
Avaluació de les millores (2)

- 40 / 100 Gigabit Ethernet
  - Major velocitat (40 / 100 Gbps) / Menor latència (0,3 µs)



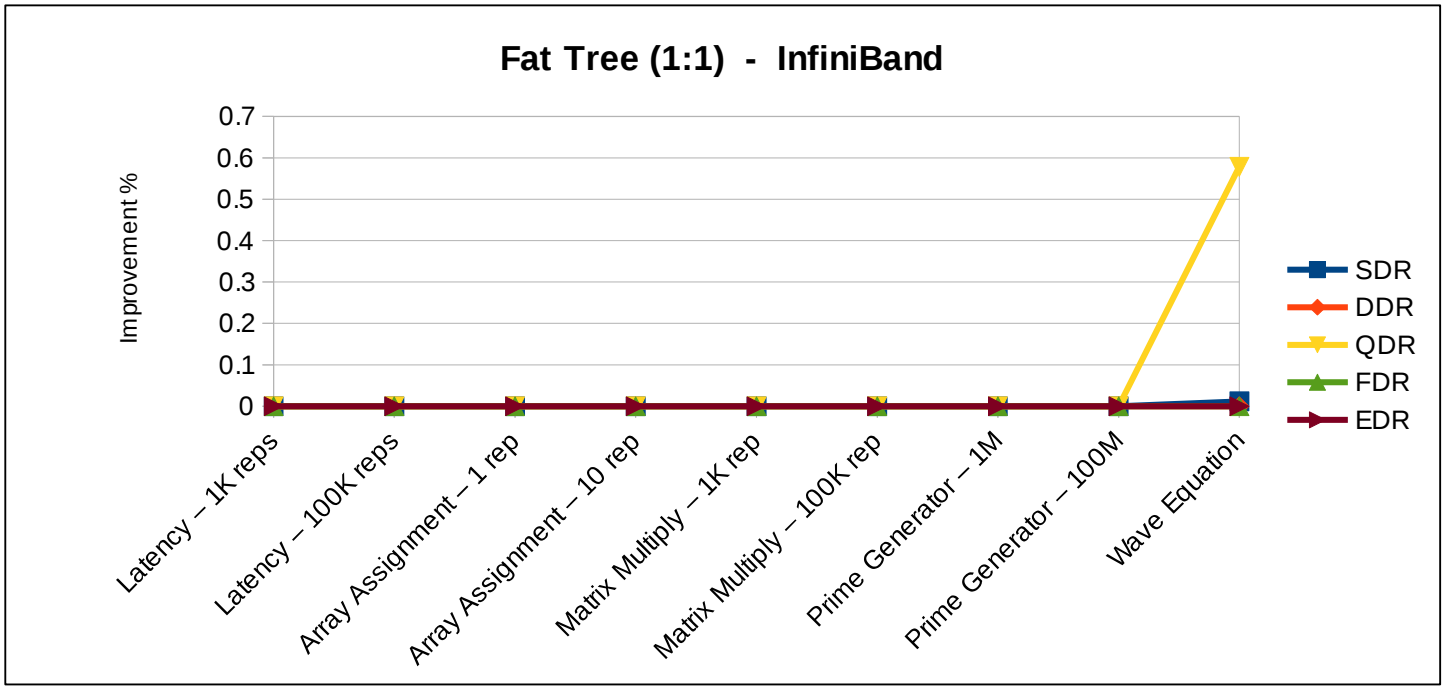
Avaluació de les millores (3)

- InfiniBand NDR
  - Major velocitat (180 Gbps) / Menor latència (0,35 µs)



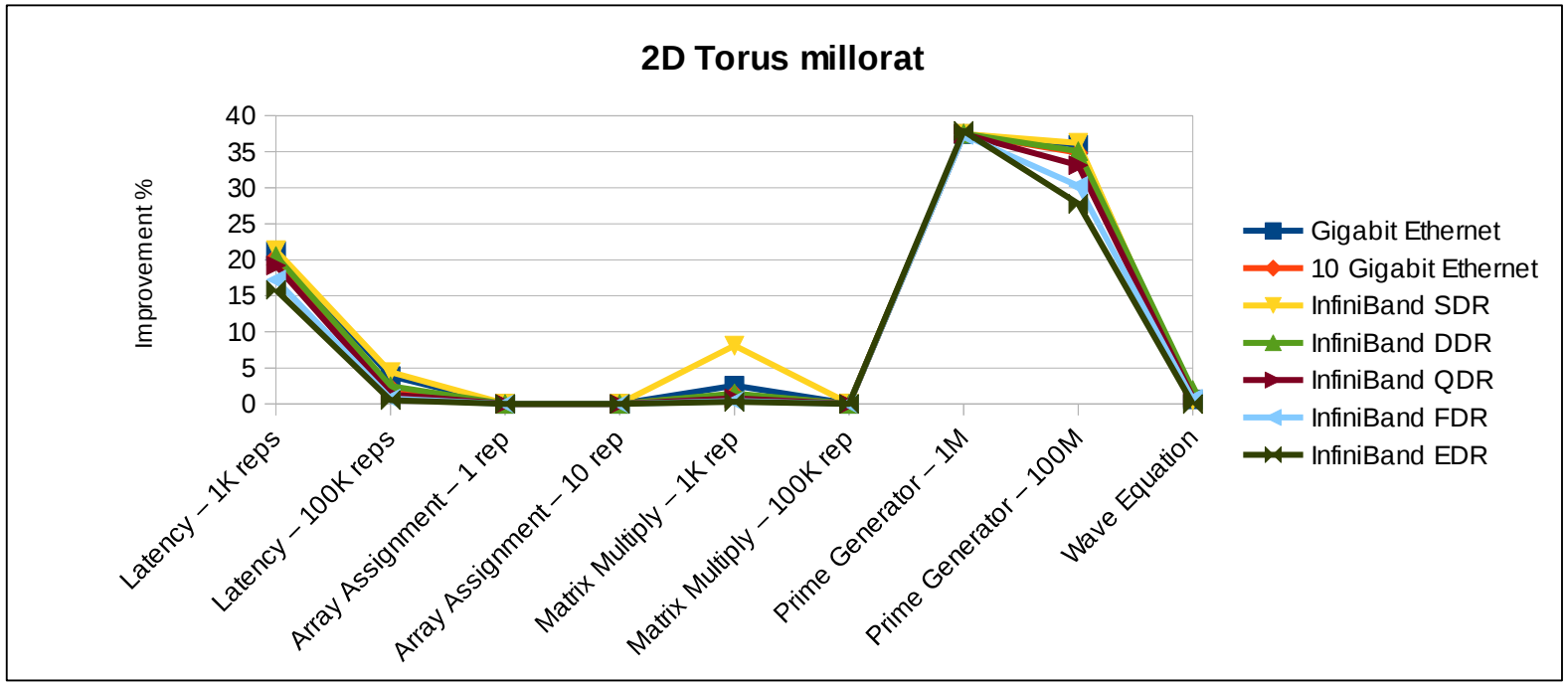
Avaluació de les millores (4)

- Fat Tree (1:1) amb InfiniBand
  - Equilibri perfecte entre nivells (relació 1:1)



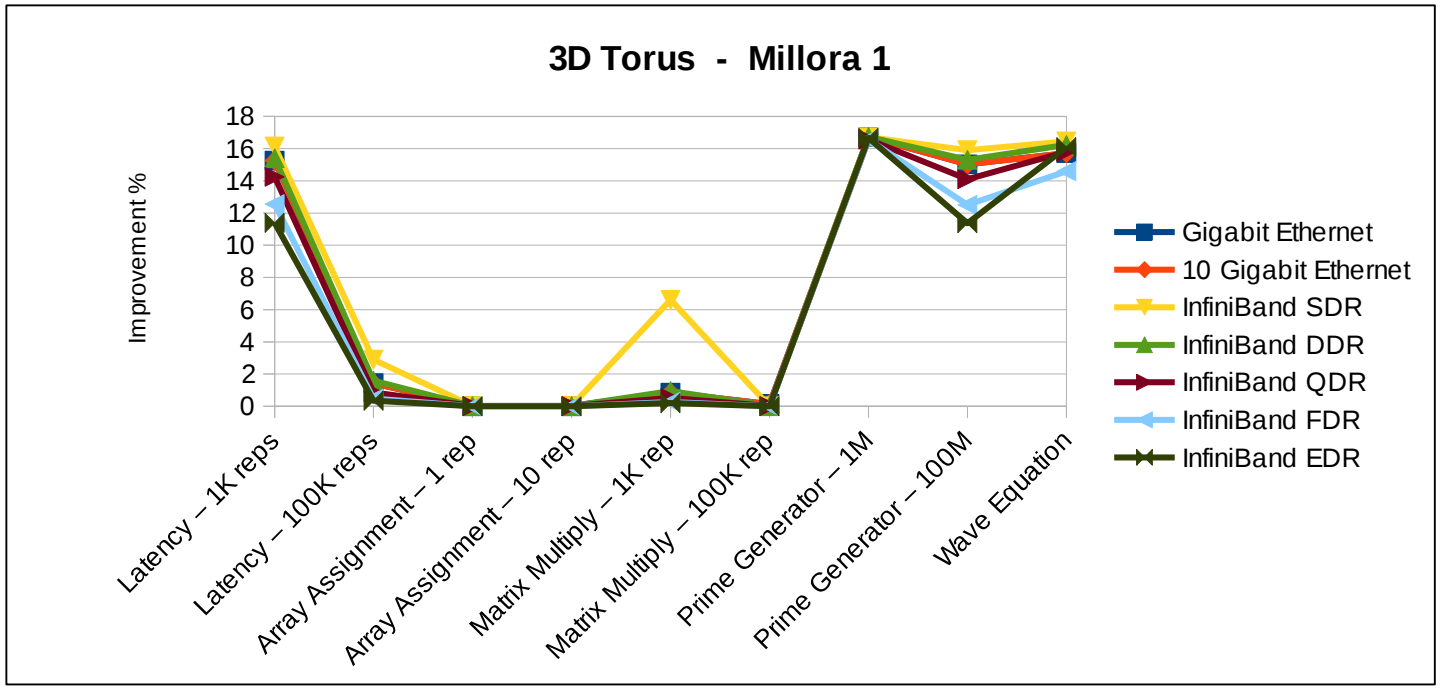
Avaluació de les millores (5)

- 2D Torus millorat
  - Augment nº enllaços (128 existents + 98 nous)



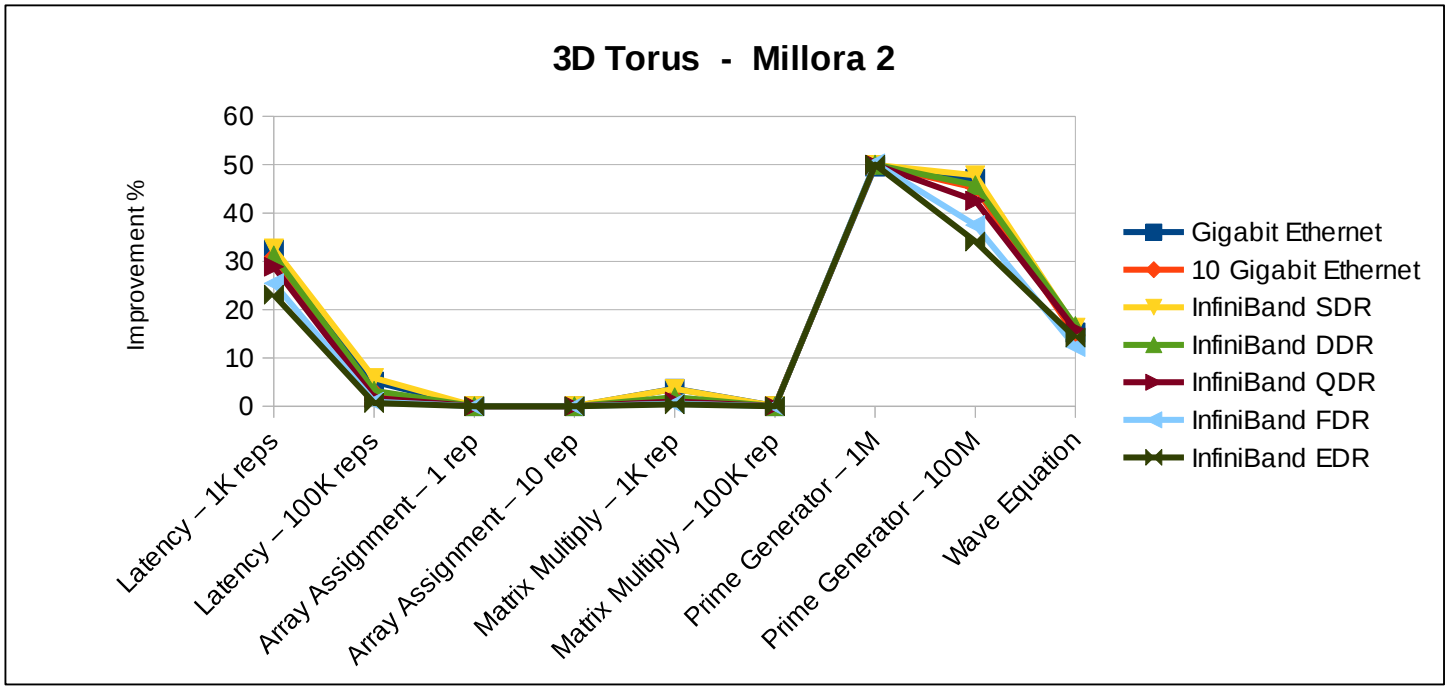
Avaluació de les millores (6)

- 3D Torus millorat
  - Millora 1: Augment nº enllaços (176 existents + 54 nous)



Avaluació de les millores (7)

- 3D Torus millorat (2)
  - Millora 2: Augment nº enllaços (176 + 54 + 16 nous)



Conclusions

- 10 Gigabit Ethernet amb SFP+
  - Reducció de latència no millora el rendiment de forma global.
  - Només dos tests MPI es beneficien i augmenten molt el rendiment: *Prime Generator* i *Wave Equation*
  - La natura del test MPI determina si té sentit aquesta millora tecnològica.

Conclusions (2)

- 40 / 100 Gigabit Ethernet
  - Les tres topologies es beneficien d'aquesta millora.
  - Fat Tree és la topologia que més es beneficia.
  - *Prime Generator* i *Wave Equation* són els tests MPI que més augmenten el rendiment. La resta també, però en menor mesura.
  - Podria ser interessant aplicar aquesta evolució en un entorn d'alt rendiment, per tal de millorar els temps d'execució.

Conclusions (3)

- Infiniband NDR
  - Fat Tree augmenta el seu rendiment clarament.
  - Les topologies Torus només amb dos dels tests MPI: *Prime Generator* i *Wave Equation.*
  - La natura del test MPI i la topologia utilitzada determinen si aquesta millora tecnològica es veu justificada.
  - En cas de tindre la topologia Fat Tree, té justificació evolucionar cap aquesta tecnologia.

Conclusions (4)

- Fat Tree 1:1 amb InfiniBand
  - No es justifica per als tests MPI fets servir.
  - Connexions troncals (entre switches) amb agregacions tipus 12x suficients per a les simulacions utilitzades.

Conclusions (5)

- 2D Torus millorat
  - De forma general, augmentar el nombre de connexions no millora en rendiment.
  - Només un test MPI millora els temps d'execució: *Prime Generator*.
  - La natura del test MPI determina si té sentit aplicar aquesta millora.

Conclusions (6)

- 3D Torus millorat
  - De forma general, augmentar el nombre de connexions no millora en rendiment.
  - Només dos tests MPI milloren els temps d'execució: *Prime Generator* i *Wave Equation.*
  - La natura del test MPI determina si té sentit aplicar aquesta millora.

Treball futur

- Tecnologies d'interconnexió
  - Rendiment de tecnologies obsoletes: Token Ring, Myrinet
  - Rendiment de tecnologies futures: 400 Gigabit Ethernet
- Topologies
  - Fat Tree: Augmentar/Reduir nombre switches Edge/Core
  - 2D Torus: Augmentar/Desequilibrar la malla
  - 3D Torus: Augmentar/Desequilibrar el cub
  - Hypercub: 4D Torus , 5D Torus
- Altres
  - Programari MPI: definir altres tests
  - Augmentar nombre de nodes
  - Major rendiment càlcul per node (FLOPS)
  - Altres tecnologies de processament: GPGPU

Fi de la presentació

José Antonio Martín Pérez
jmartinperez1@uoc.edu