# Cray XT4 with Quad-core Processors: A First Look

Sadaf R. Alam, Richard F. Barrett, Markus Eisenbach, Mark R. Fahey,
Rebecca Hartman-Baker, Jeffrey A. Kuehn, Stephen W. Poole, Ramanan Sankaran

Scientific Computing Group

National Center for Computational Sciences
Oak Ridge National Laboratory

http://nccs.gov

and Patrick Worley
Computer Science and Math Division

Cray User Group 2008, Helsinki
May 6, 2008

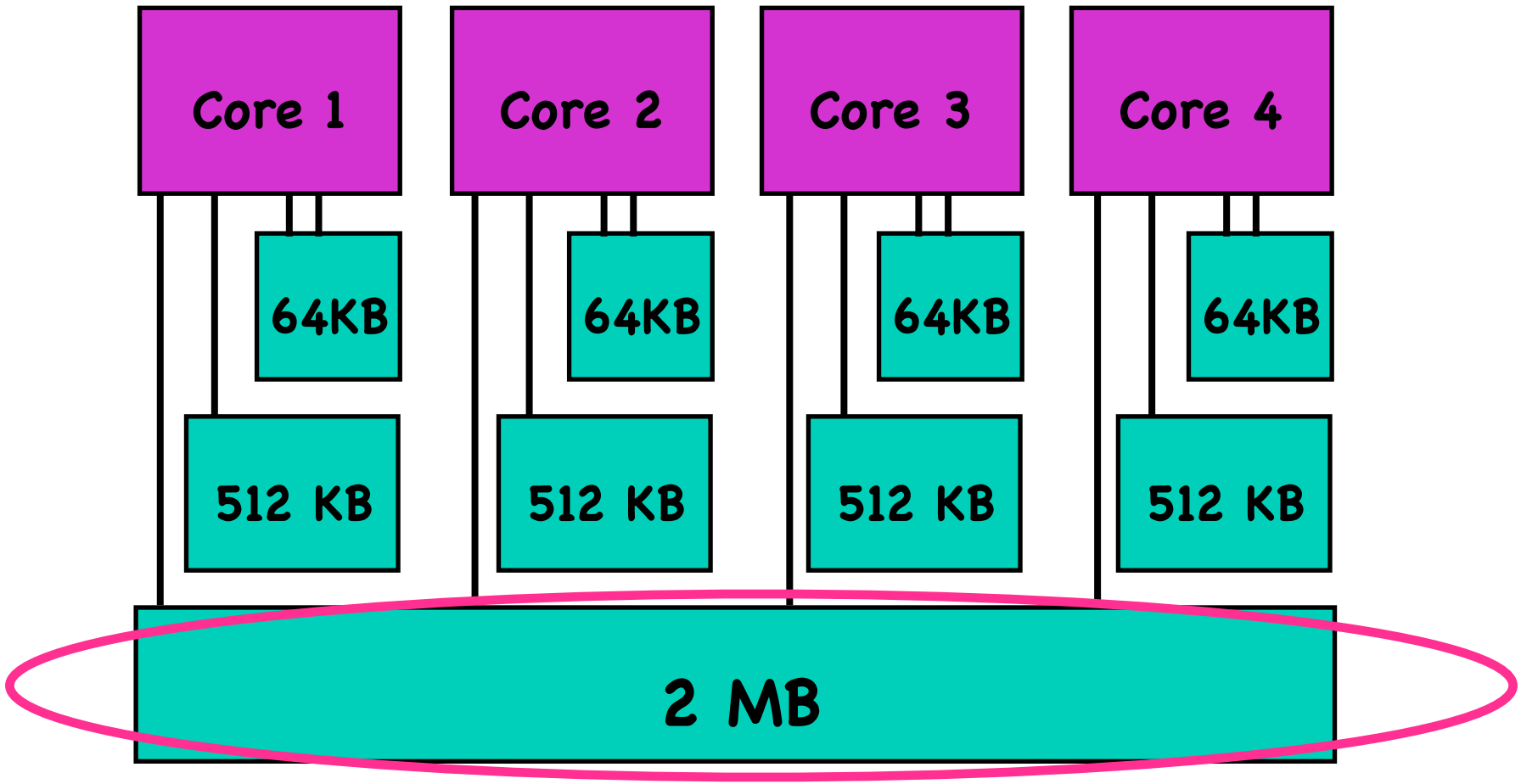
OAK RIDGE National Laboratory

# XT@ORNL History : Speeds and Feeds

| System | | | | Processor Speeds | | | Main Memory | | | Network (3d torus) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Arch | Date | cores /node | Num Nodes | GHz | Ops/ clock | Peak TFLOPS | GB/ node | DDR | Mem GB/s | Sea Star | BW GB/s |
| XT3 | 6/05 | 1 | 5,212 | 2.4 | 2 | 25 | 2 | 1 | 6.4 | 1 | 2.2 |
| XT3 | 7/06 | 2 | 5,212 | 2.6 | 2 | 54 | 4 | 1 | 6.4 | 1 | 2.2 |
| XT3/4 | 4/07 | 2 | 11,508 | 2.6 | 2 | 119 | 4 | 1,2 | 6.4/10.6 | 1,2 | 2.2,4 |
| XT4 | 5/08 | 4 | 7,832 | 2.1 | 4 | 263 | 8 | 2 | 10.6 | 2 | 4 |

"An Evaluation of the ORNL Cray XT3", Alam, et al, IJHPCA 22(1), Feb 2008. (Single core)
"Cray XT4: An Early Evaluation for Petascale Scientific Simulation", Alam, et al. SC'07. (DC)
"Cray XT4 Quad-core: A First Look", Alam, et al, CUG 2008. (QC)

# XT4 Quad-core Barcelona processor



# Software

Compute Node Linux (CNL) lightweight kernel

Portals data movement layer

PGI, PathScale, gnu.

Cray libsci, AMD Core Math Library (ACML)

Lustre parallel file system

# What we'll see :

Bad news: simple port will probably be disappointing.

Good news:

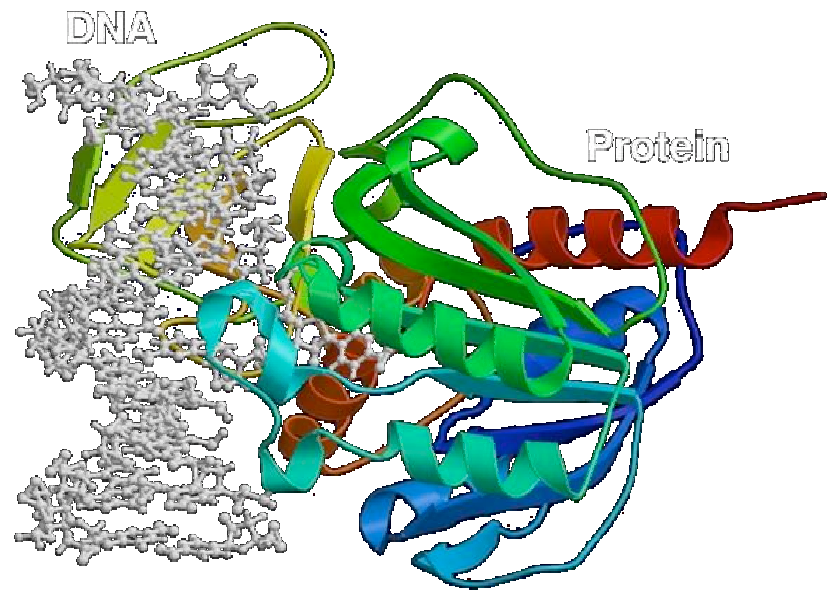
Significant improvements with little effort
Significant improvements with some effort
Significant improvements with bigger effort

Resulting in stronger code.

# Molecular Dynamics: LAMMPS

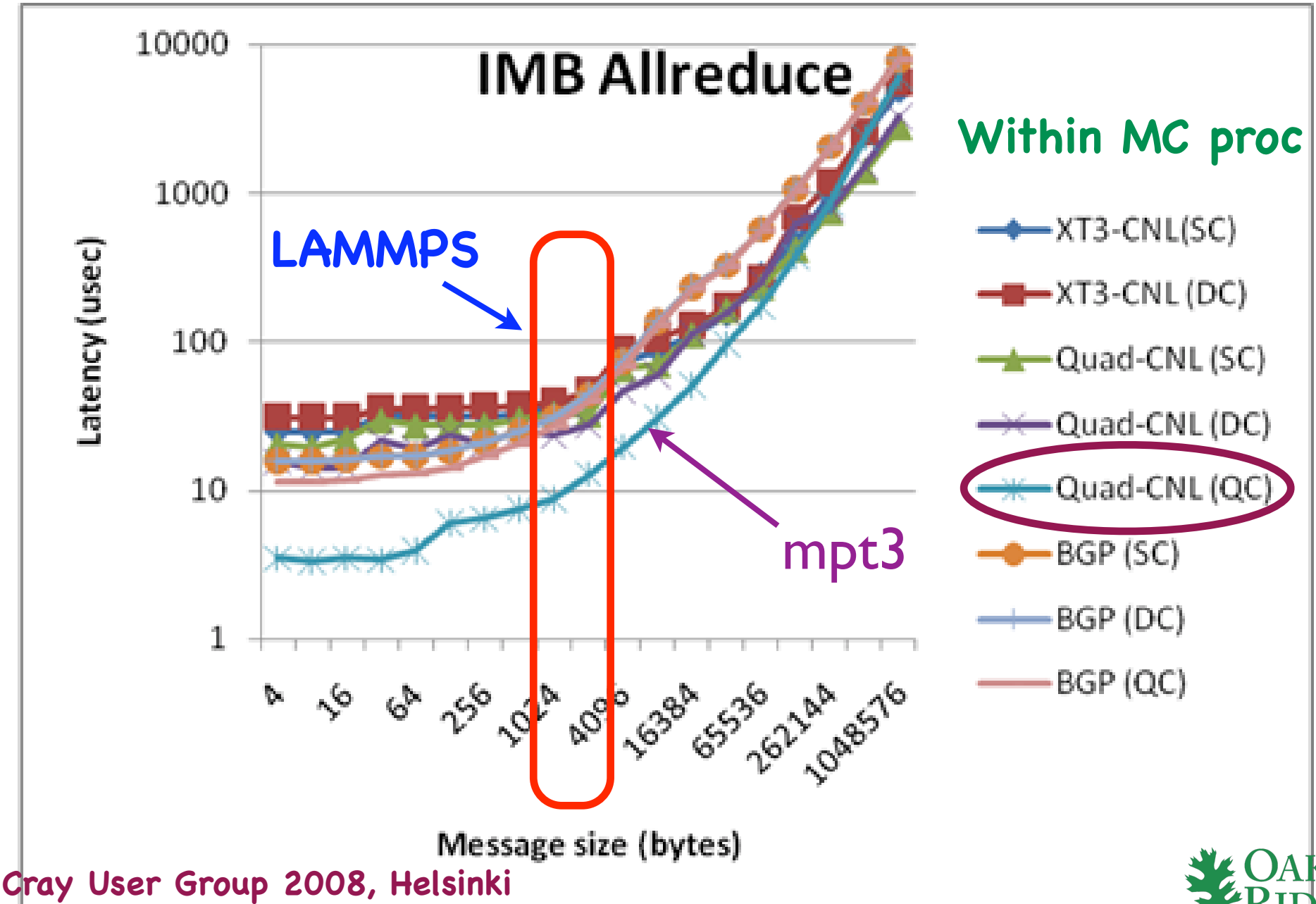
Large-scale Atomic/Molecular Massively Parallel Simulator

- models an ensemble of particles in a liquid, solid, or gas
  - atomic, polymeric, biological, metallic or granular sys.
- C++ and MPI.

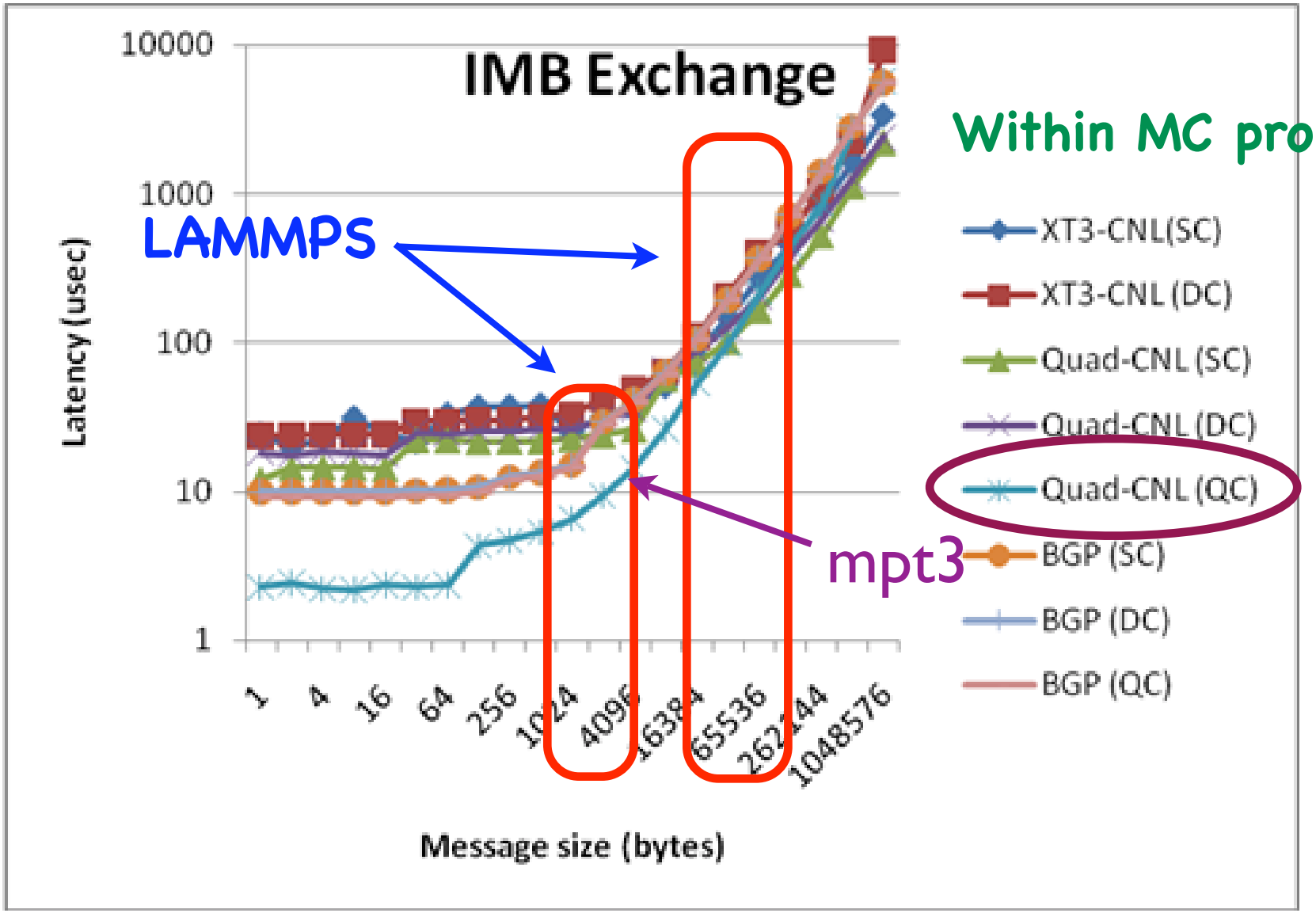



*Protein-DNA complex for the M.HhaI methyltransferase*

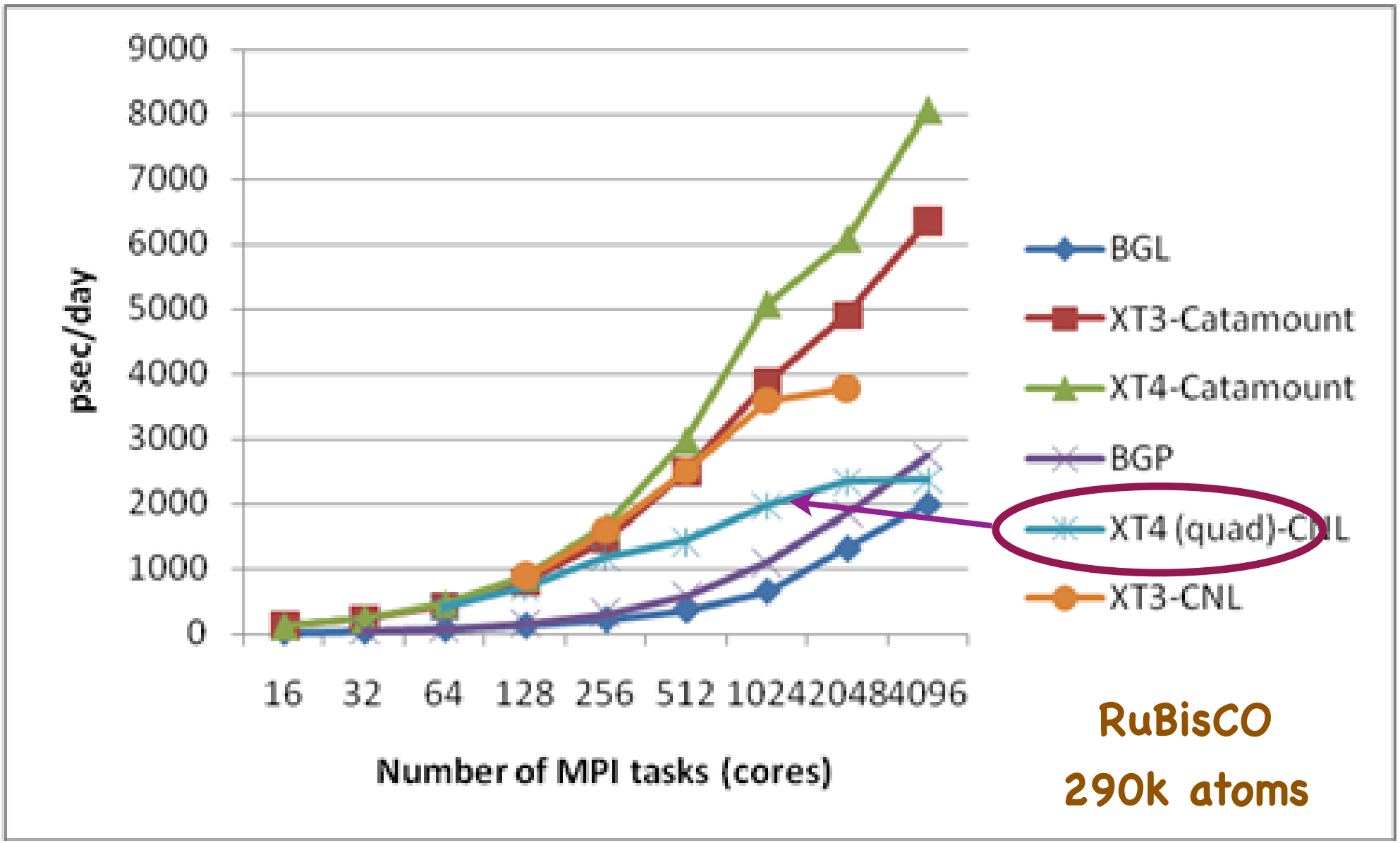
# LAMMPS key communication ( MPI ) :



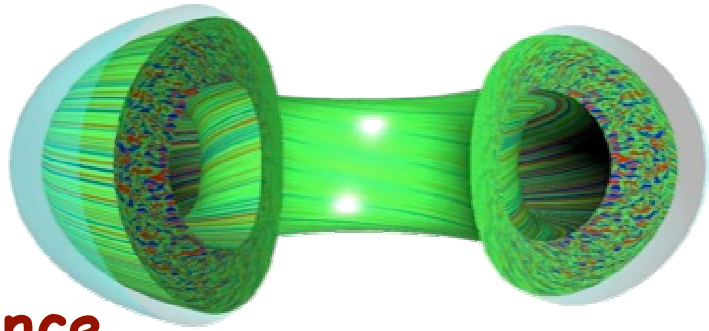
# LAMMPS key comm ( MPI ) :



# LAMMPS performance

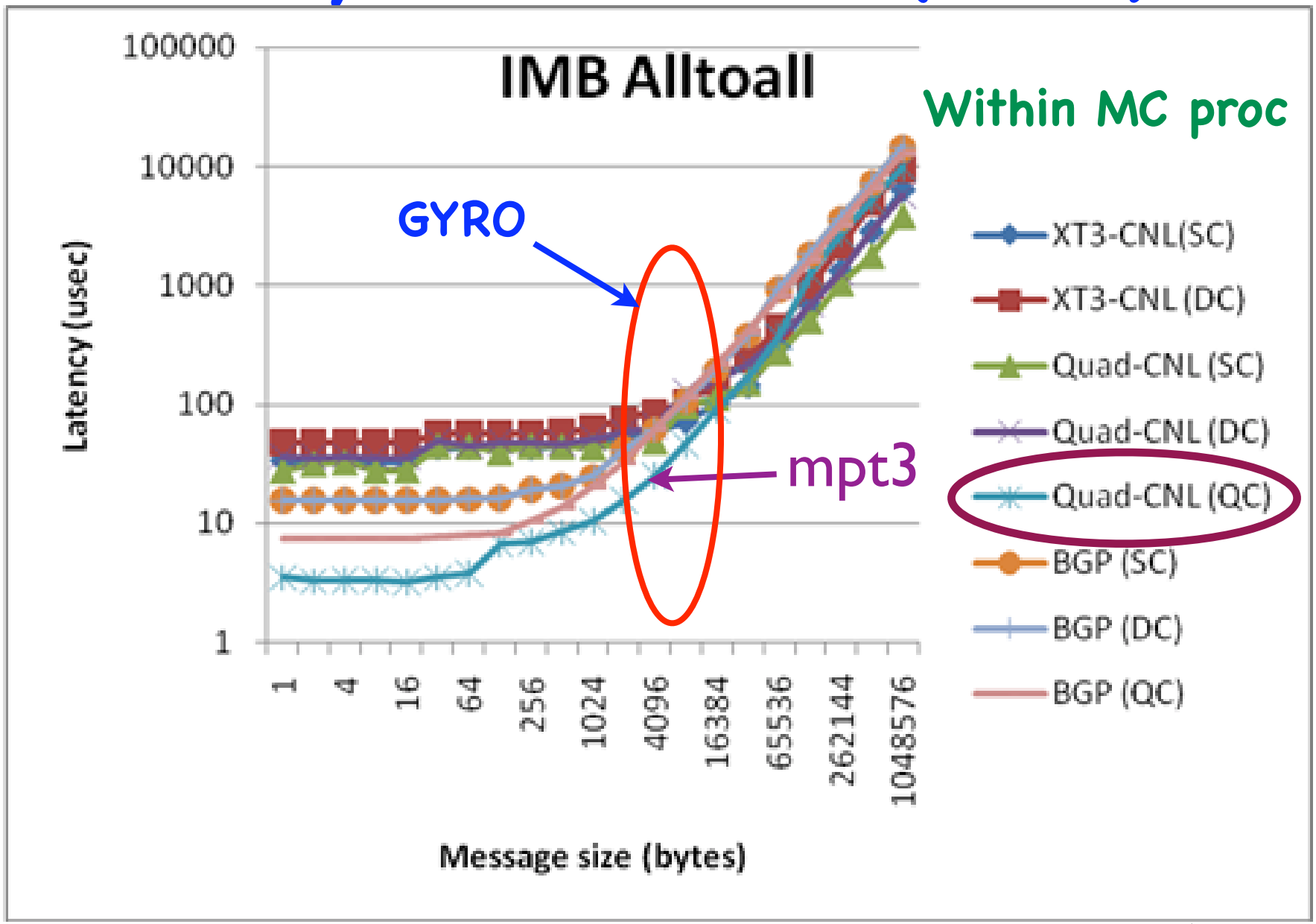


# Fusion Energy : GYRO

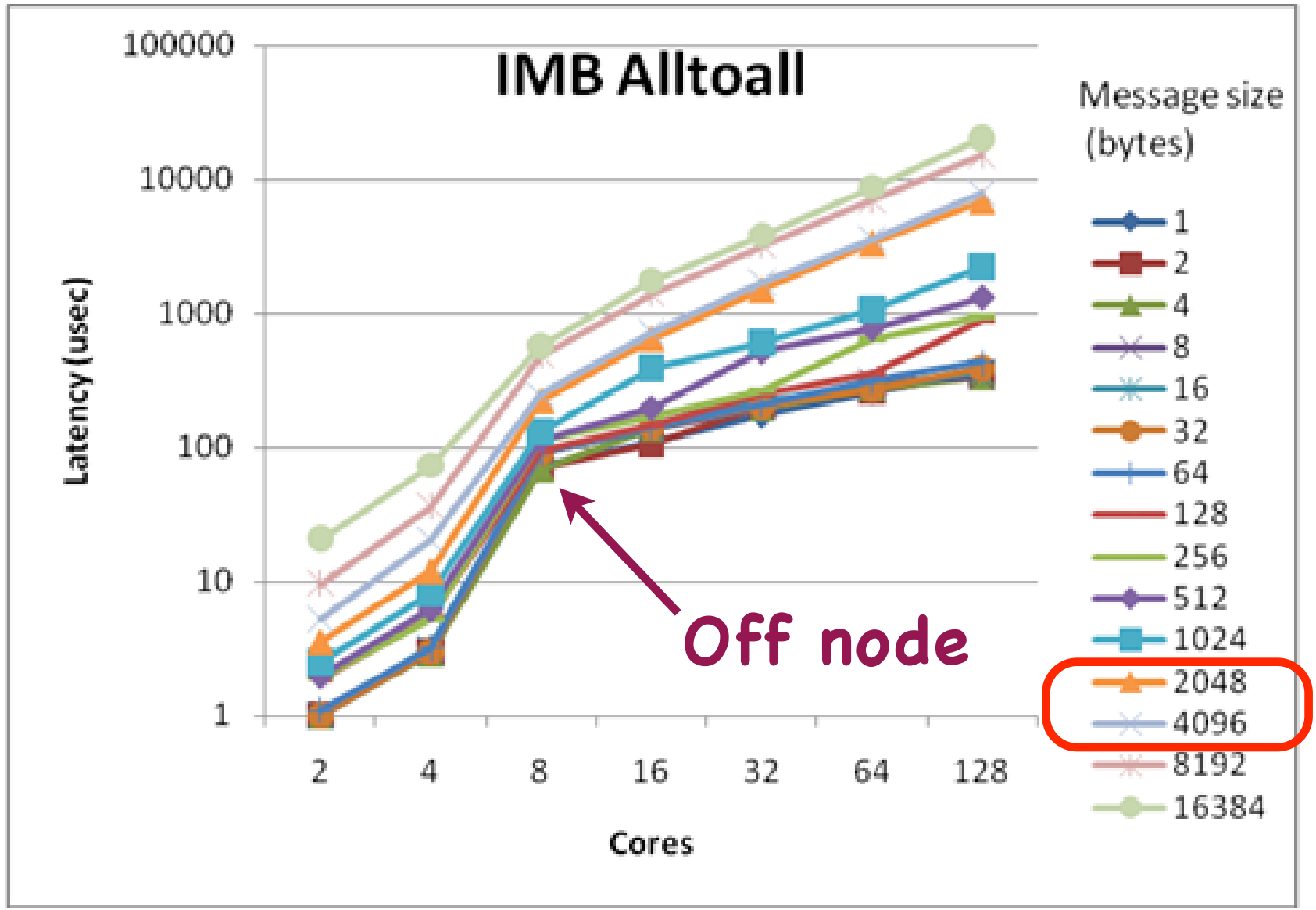


- simulation of tokamak microturbulence
- time-dependent, nonlinear gyrokinetic-Maxwell equations
- five-dimensional grid
- fourth-order, explicit, Eulerian algorithm.
- INCITE: "Gyrokinetic steady state transport simulations"

# GYRO key communication ( MPI ) :

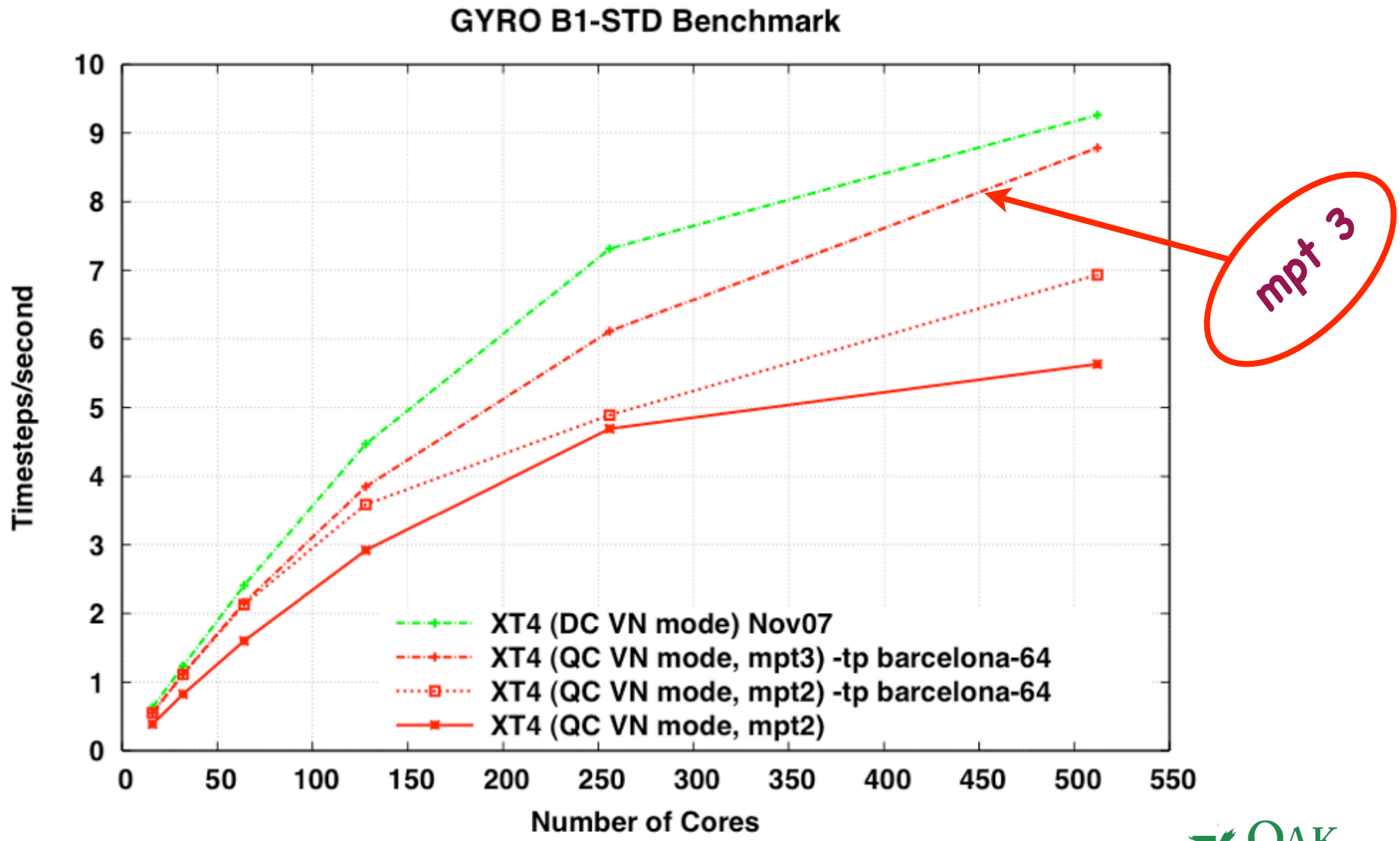


# GYRO key communication ( MPI ) :



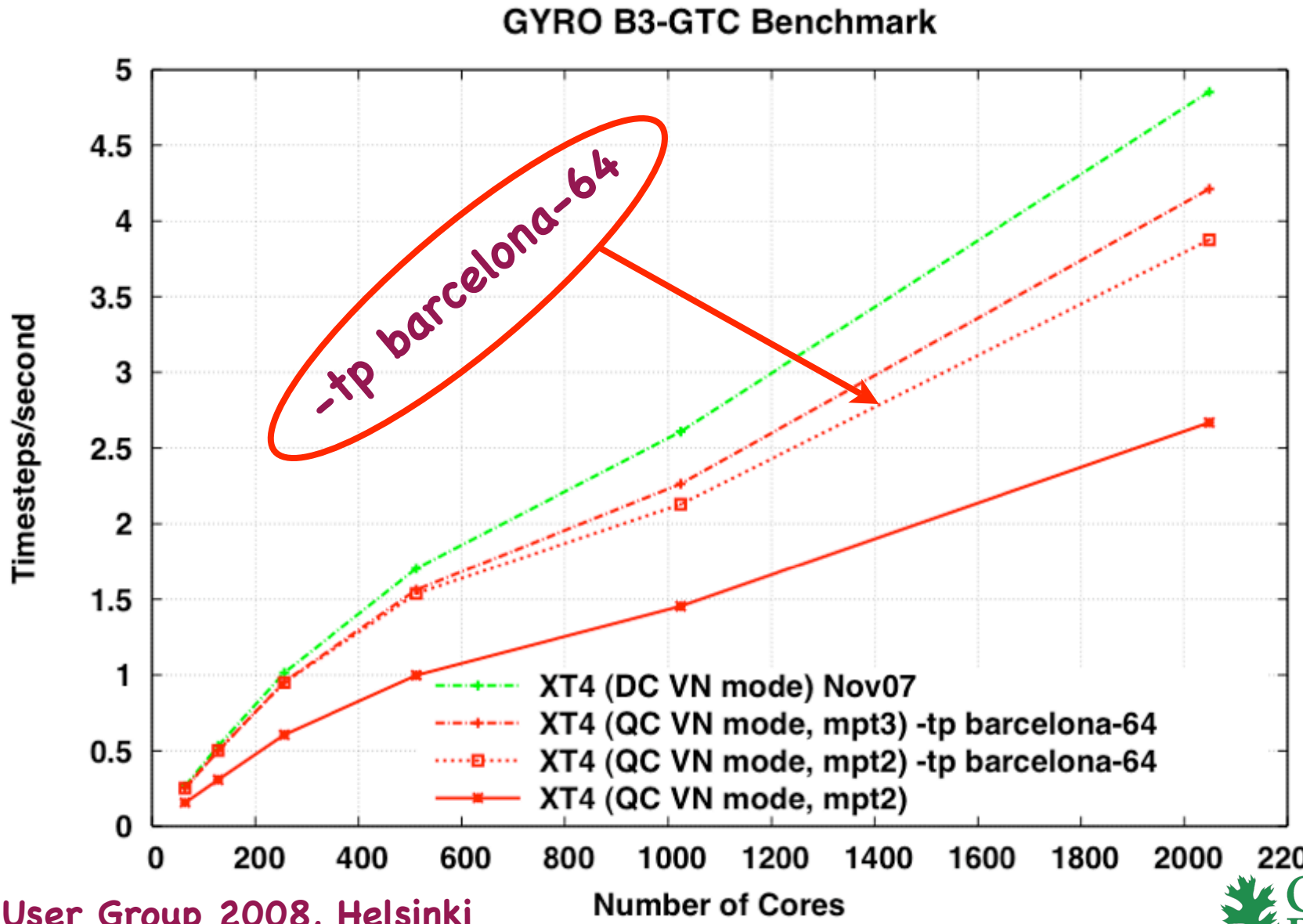
# GYRO performance : B1

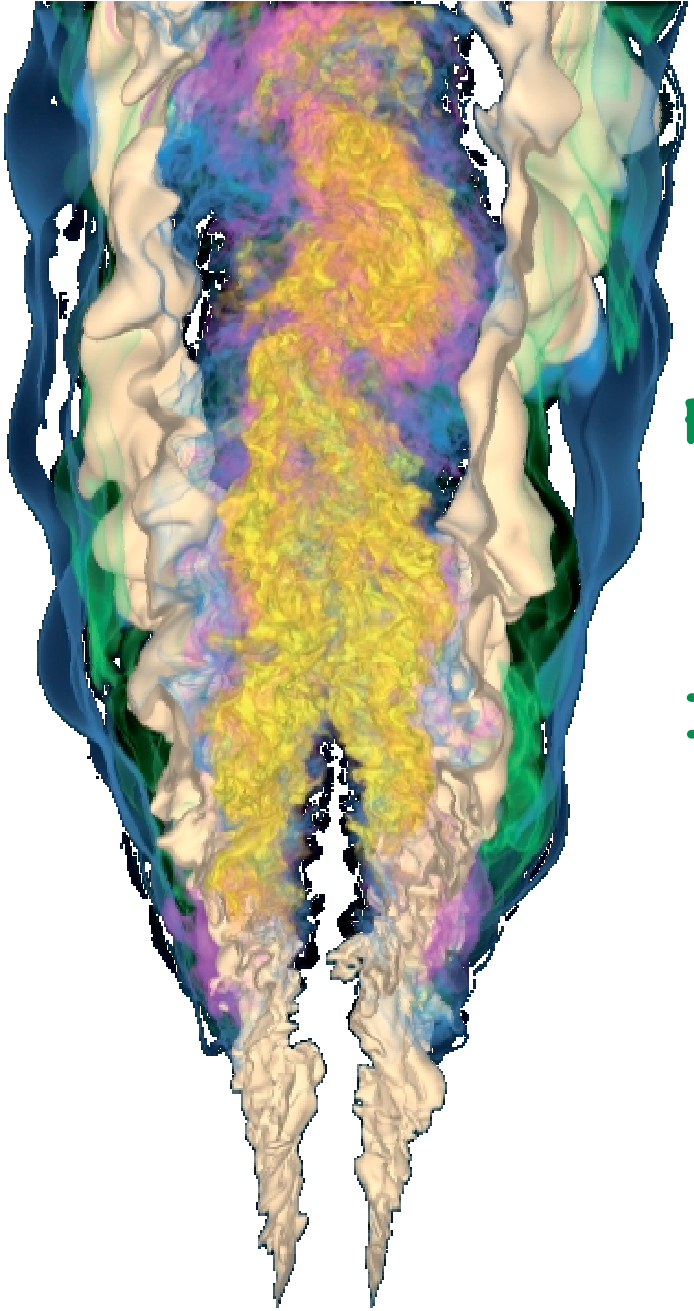
- 16 x 140 x 8 x 8 x 20 grid, 100 timesteps (3 sim secs)



# GYRO performance : B3

## 64 x 400 x 8 x 8 x 20 grid, 100 timesteps (3 sim secs)





# Combustion turbulence: S3D

Flow solver for performing direct numerical simulation (DNS).

INCITE: "High-Fidelity Simulations for Clean Efficient Combustion of Alternative Fuels"

*DNS of a turbulent lifted jet flame performed by Chun Sang Yoo and Jacqueline Chen, Sandia National Labs. The image was rendered by Hongfeng Yu and Kwan-Liu Ma of UC Davis and SciDAC Ultrascale Visualization Institute.*

OAK RIDGE National Laboratory

# S3D single processor performance

| Problem size | MPI mode (aprun) | Dual-core | | Quad-core | |
|---|---|---|---|---|---|
| | | wall time (secs) | usec / gridpt / tstep | wall time (sec) | usec/ grid-pt/ tstep |
| 30x30x30 | -n 1 -N 1 | 404 | 150 | 415 | 154 |
| 60x30x30 | -n 2 -N 2 | 465 | 172 | 430 | 159 |
| 60x60x30 | -n 4 -N 4 | - | - | 503 | 186 |

$C_2H_4$ : weak scaling

$30^3$ grid pts / MPI process

100 time steps

# S3D performance:
# Effect of vectorization

~60% time spent computing reaction rates:

| Times from quad-core | Before | After |
|---|---|---|
| | time (secs) | |
| Rates time | 277 | 228 |
| Total time | 415 | 367 |

# S3D performance:
# Effect of vectorization

~60% time spent computing reaction rates:

| Counters (dual-core) | Before | After |
|---|---|---|
| | x $10^9$ operations | |
| add | 182 | 187 |
| Multiply | 204 | 210 |
| Add + Mult | 386 | 397 |
| Load/store | 179 | 202 |
| SSE | 91 | 212 |

# S3D performance :
# SSE is "a big deal"

| Problem size | MPI mode (aprun) | Dual-core | | Quad-core | |
|---|---|---|---|---|---|
| | | wall time (secs) | usec / gridpt / tstep | wall time (sec) | usec/ grid-pt/ tstep |
| 30x30x30 | -n 1 -N 1 | 404 | 150 | 333 (415) | 123 (154) |
| 60x30x30 | -n 2 -N 2 | 465 | 172 | 349 (430) | 129 (159) |
| 60x60x30 | -n 4 -N 4 | - | - | 422 (503) | 156 (186) |

# Fusion Energy: AORSA

rf heating in tokamak

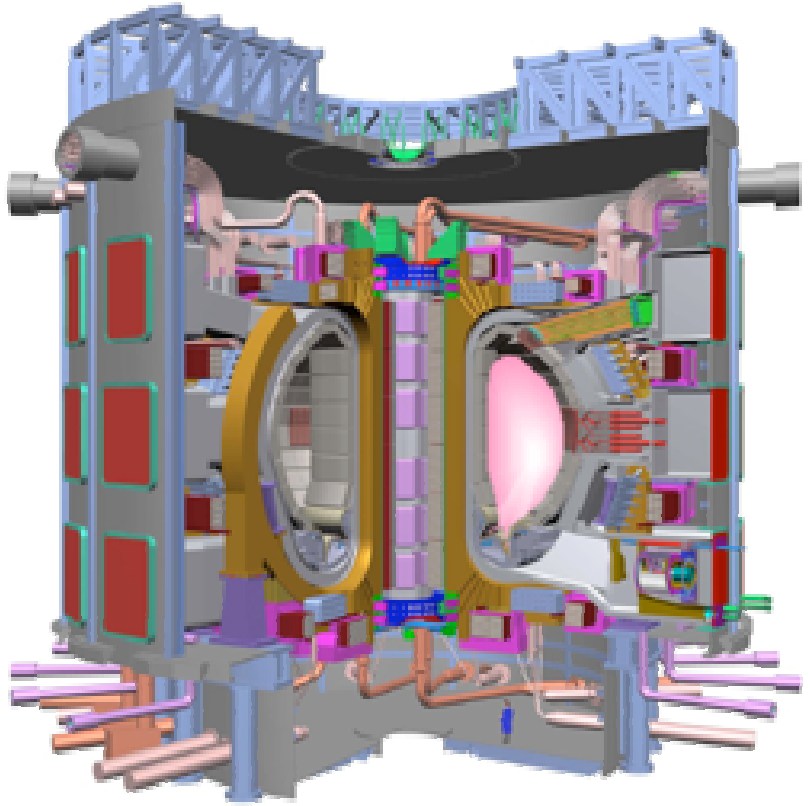
Maxwell-Bolzmann Eqns

FFT

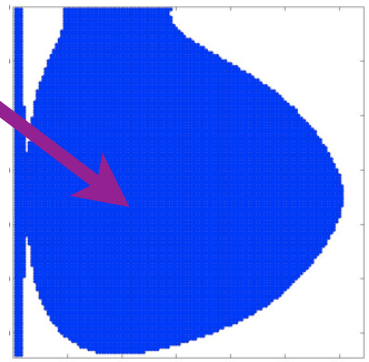
Dense linear system

Calc Quasi-linear op

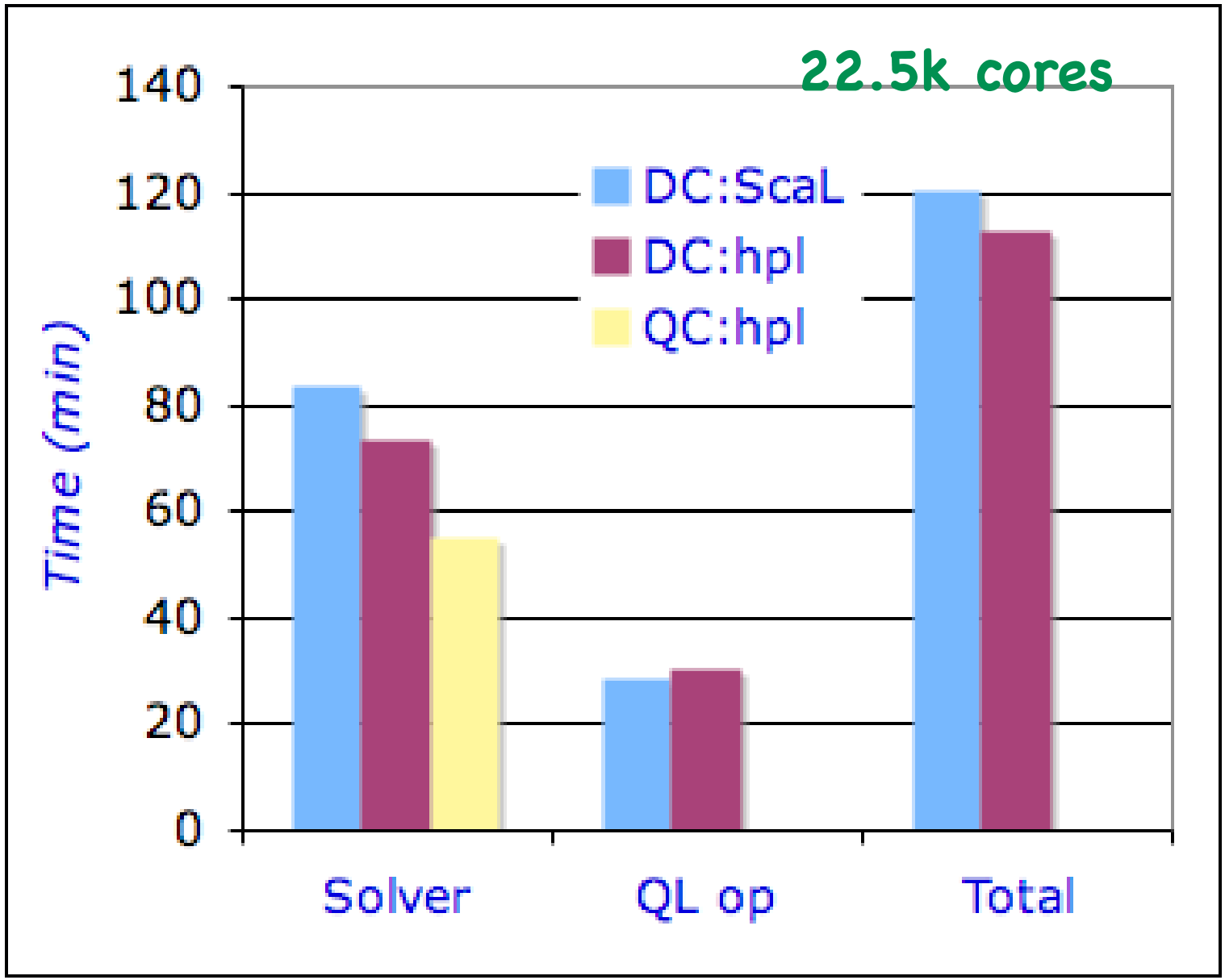
INCITE: "High Power Electromagnetic Wave Heating in the ITER Burning Plasma"



ITER-FEAT



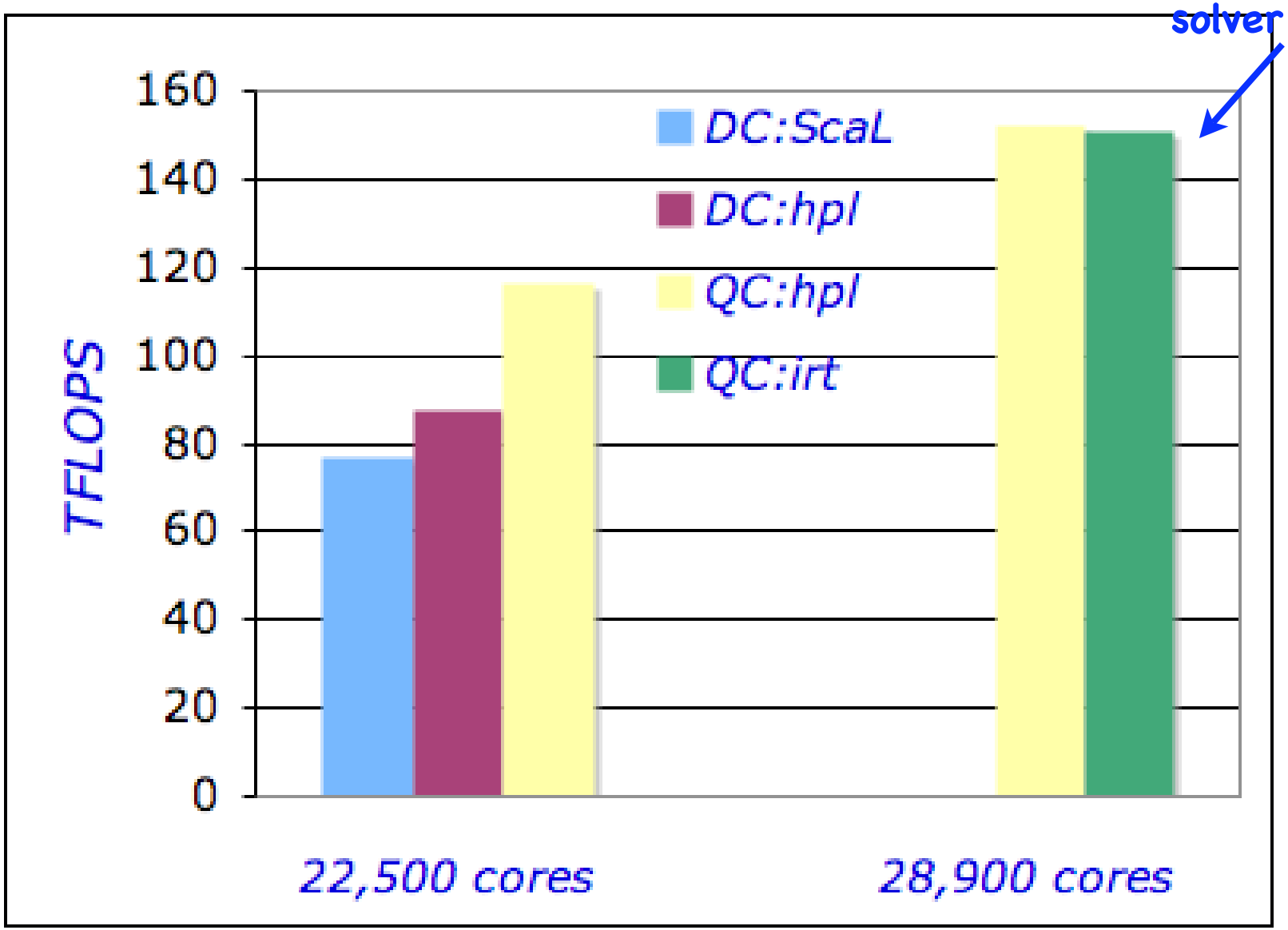
# AORSA performance




ITER on 500x500

# AORSA solver performance

Mixed precision solver in IRT

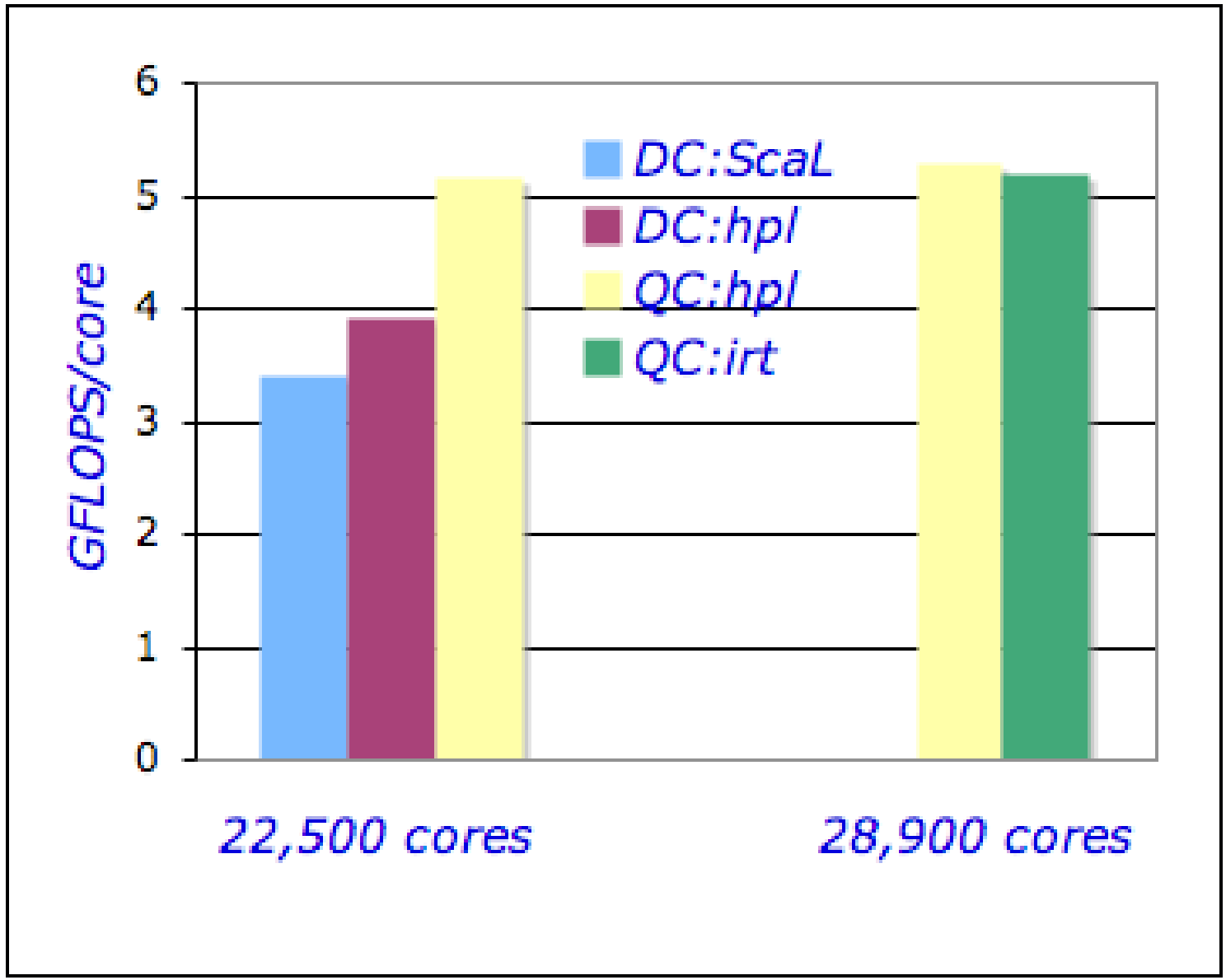



ITER on 500x500 grid

Cray User Group 2008, Helsinki
May 6, 2008

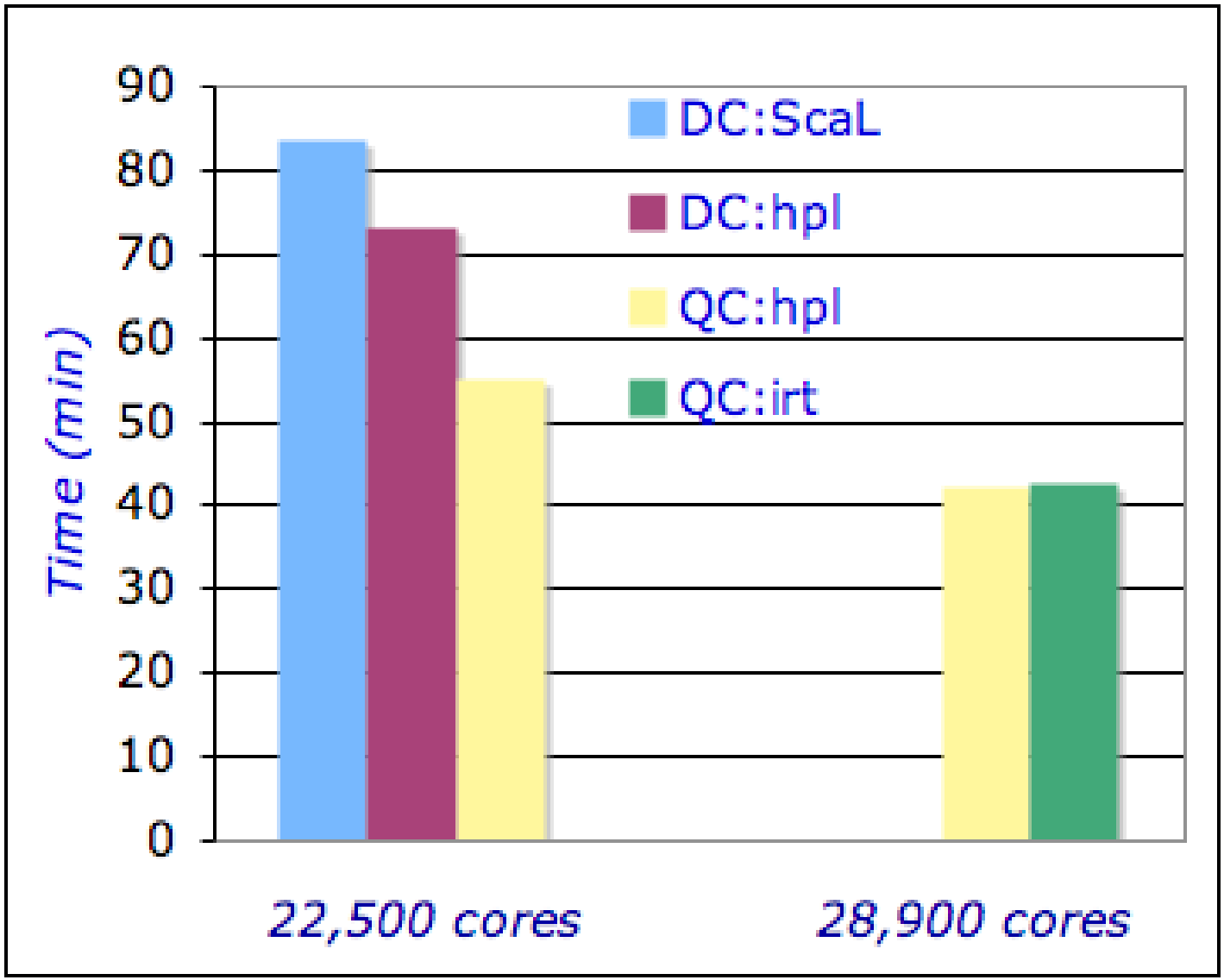
OAK RIDGE National Laboratory

# AORSA solver performance




ITER on 500x500 grid

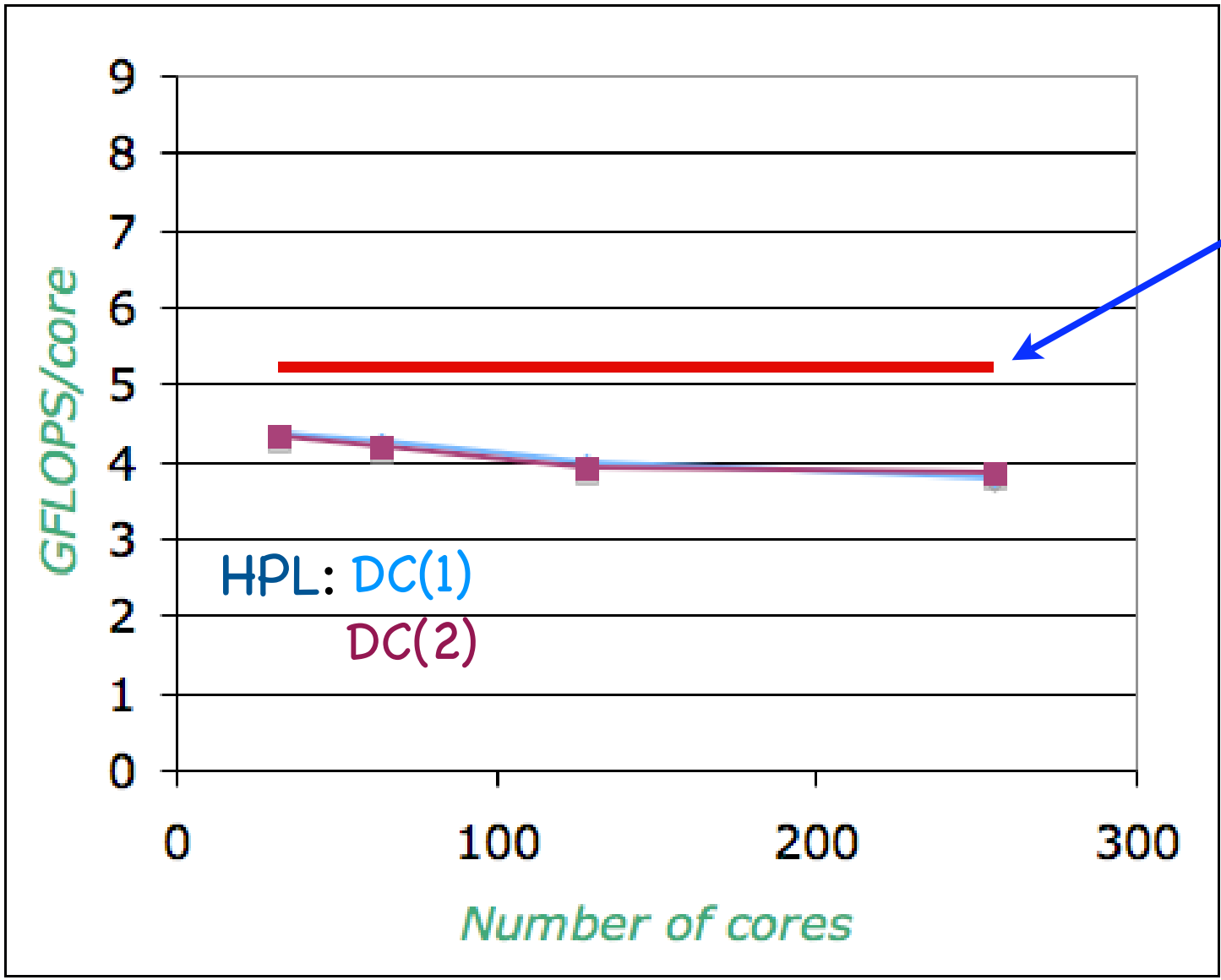
# AORSA solver performance




ITER on 500x500 grid

# AORSA solver performance

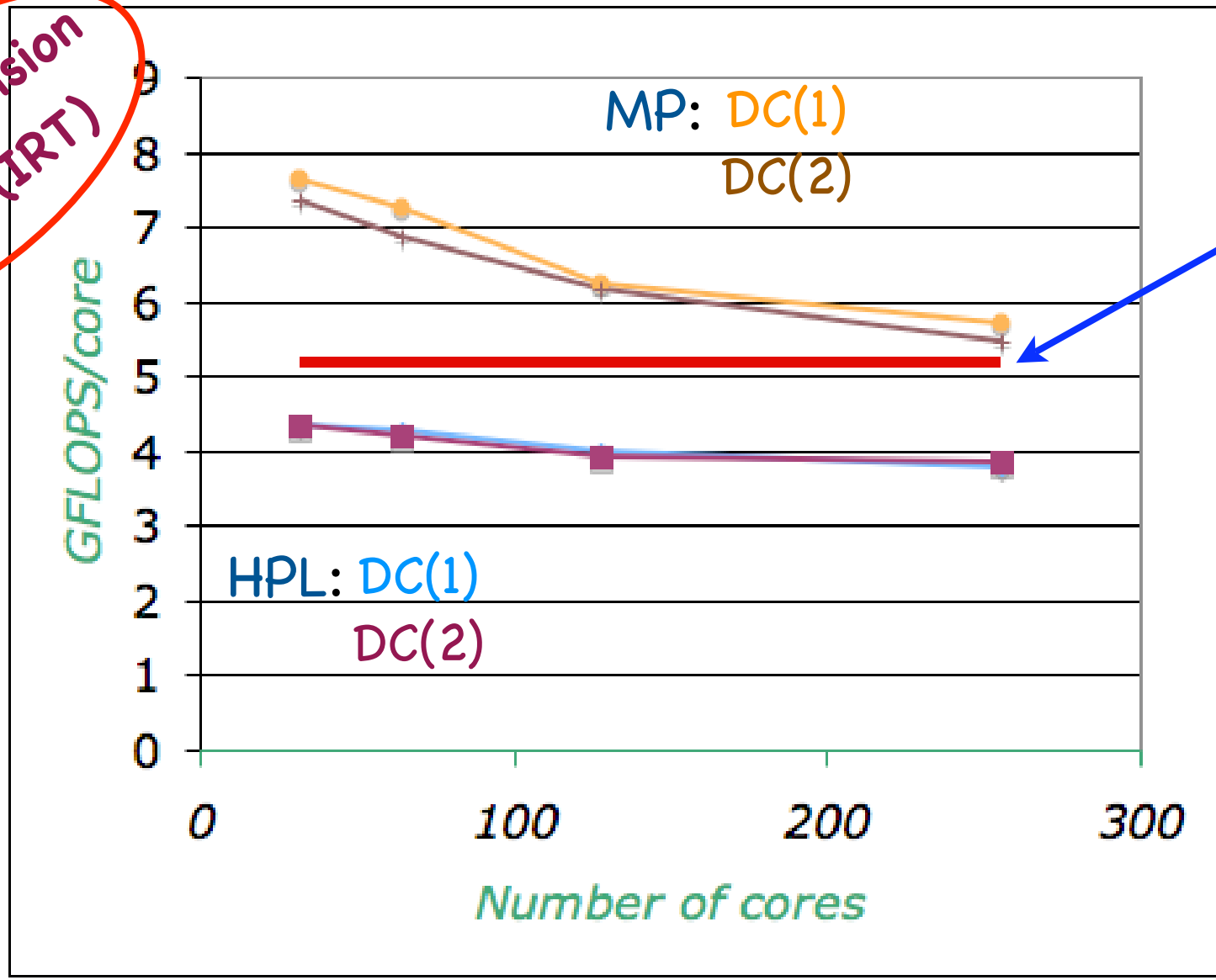
## ITER on 128x128 grid



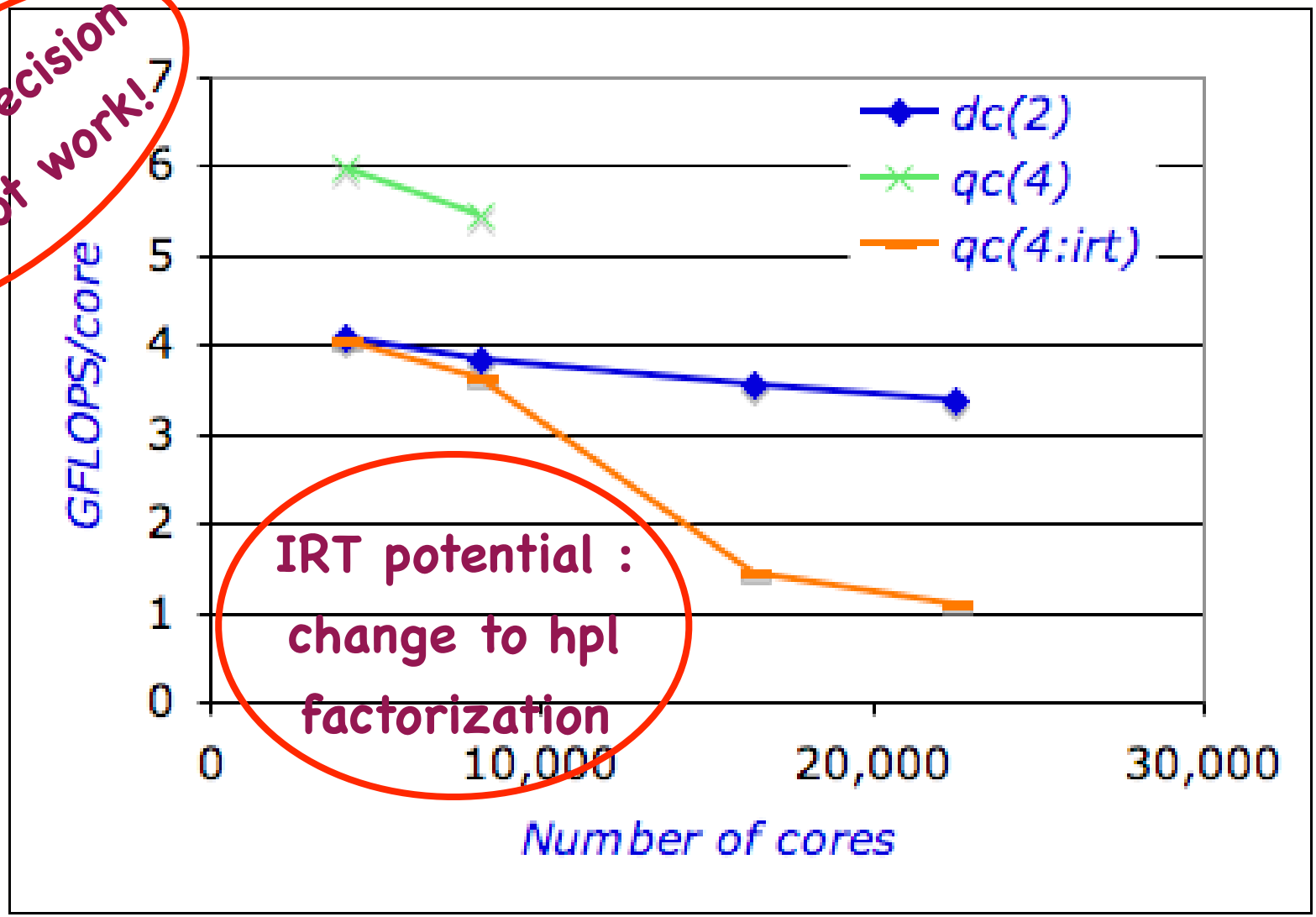
# AORSA solver performance

## ITER on 128x128 grid

Mixed-precision works! (IRT)



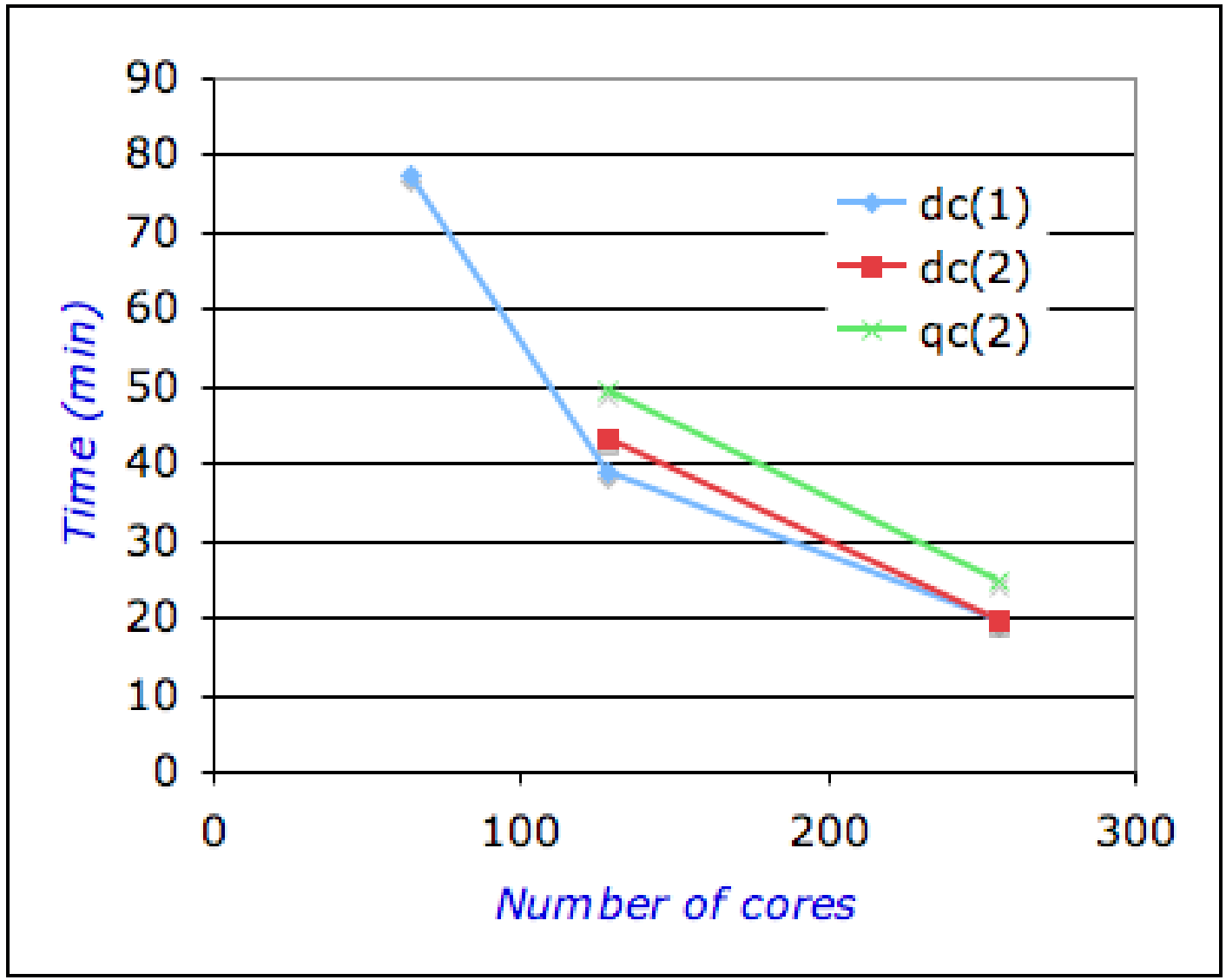
# AORSA solver performance




ITER on 350x350 grid

# AORSA key computation

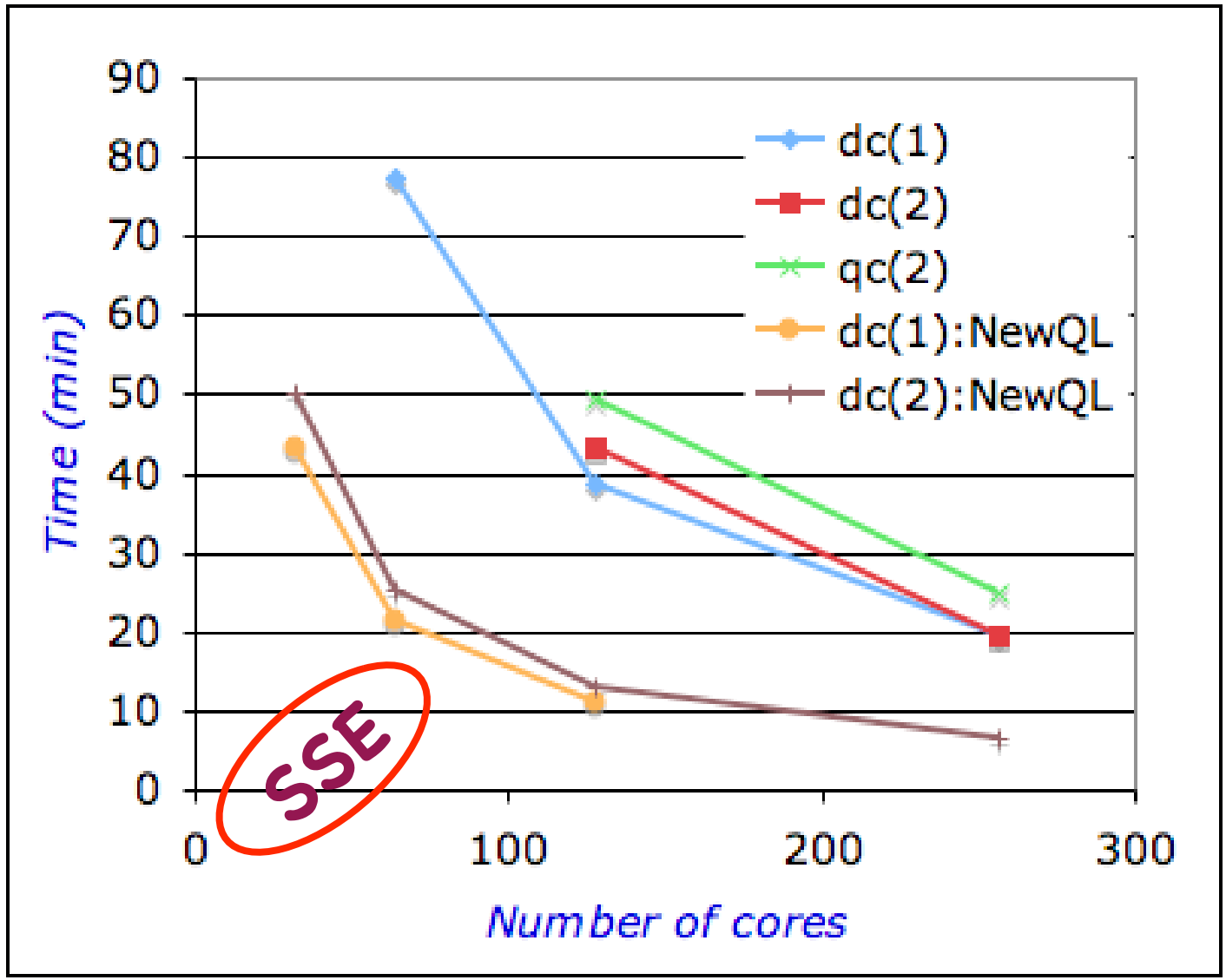
## Quasi-Linear Operator




128x128 grid

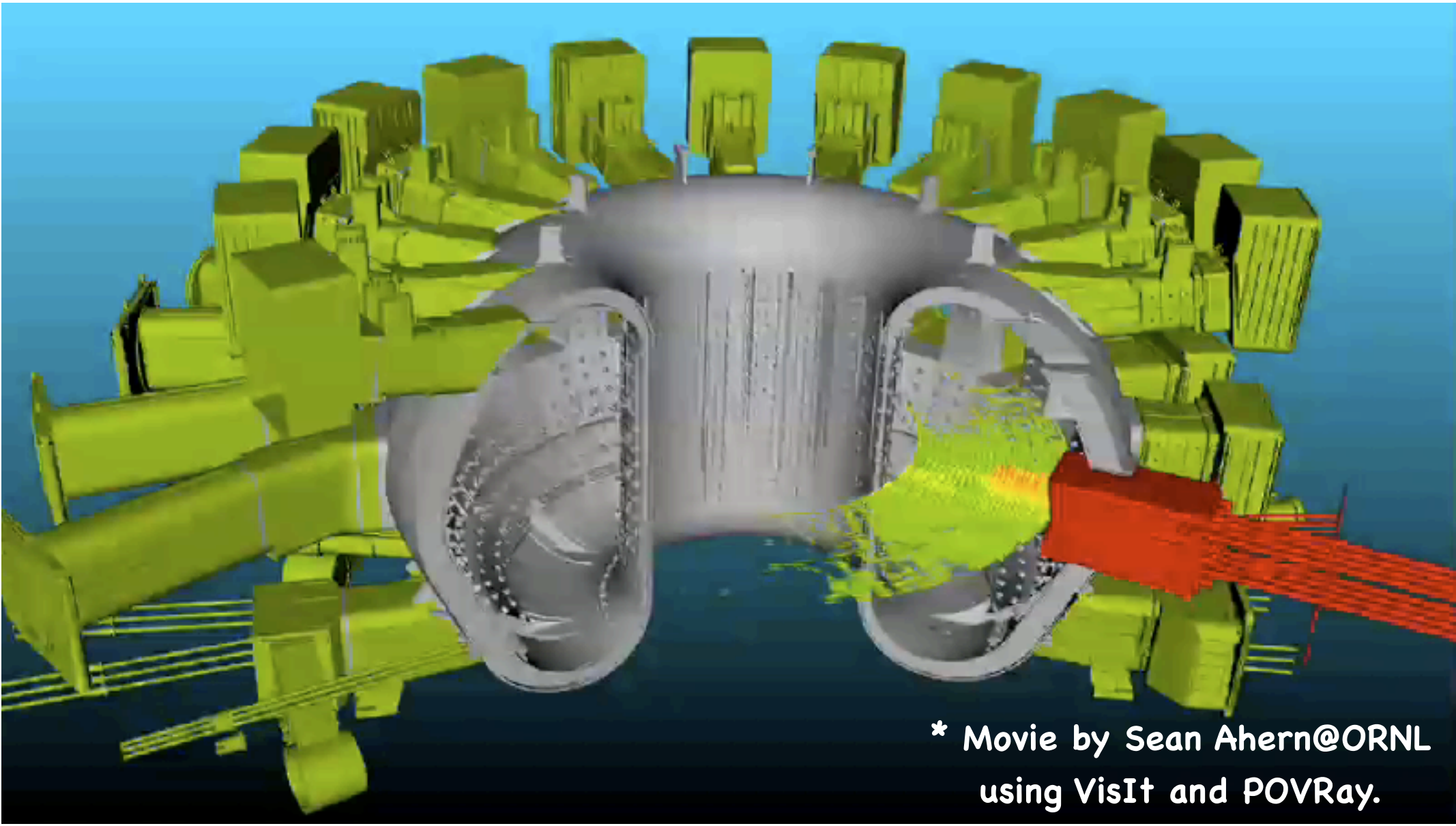
# AORSA key computation

## Quasi-Linear Operator




128x128 grid

# AORSA ITER movie*




Cray User Group 2008, Helsinki
May 6, 2008

OAK RIDGE National Laboratory

# Quad-core Performance Challenges

- It's the memory contention, right?
- Network contention?
- Will it help to make multi-core aware?
- Won't the vendors "just fix it"?
- What of the HPCS languages?

# Multi-core Performance Example3

## SPEC OMPL2001 (SPEC-HPG OpenMP benchm

## 313.swim_l (shallow water ocean model)

| Flags | Runtime in secs. (mins.) |
|---|---|
| Ofast | 7194s (120m) |
| Ofast_simd0 | 1736s (29m) |
| Ofast_movnti2500 | 1785s (30m) |

# Performance Counters

Ofast = default
Ofast_simd0 = -LNO:simd=0
Ofast_movnti2500 = -CG:movnti=2500

## Pathscale compiler

| | | default | "-LNO:simd=0" | "-CG:movnti=2500" |
|---|---|---|---|---|
| | CPI | 14 | 2.55 | 3.7 |
| | Clocks(B) | 15373 | 4107 | 4146 |
| | Insts(B) | 1100 | 1618 | 1096 |
| | L3Req | 440.000 | 161.800 | 142.480 |
| | L3Miss | 330.000 | 134.294 | 134.808 |
| | totSSE | 792.000 | 1164.960 | 789.120 |
| absolute(B) | FPadd pipe | 583.000 | 388.320 | 252.080 |
| absolute(B) | FPmult pipe | 242.000 | 210.340 | 151.248 |
| absolute(B) | FPstore pipe | 291.500 | 142.384 | 113.984 |
| | PgOpen | 253.000 | 37.214 | 36.168 |
| | PgClose | 233.200 | 114.878 | 111.792 |
| | PgCflct | 114.4 | | |

* Pathscale ™ Compiler Suite, Version 3.1, SLES10 SP1

Slide courtesy of Brian Waldecker@AMD, NCCS Quadcore workshop, April 2008

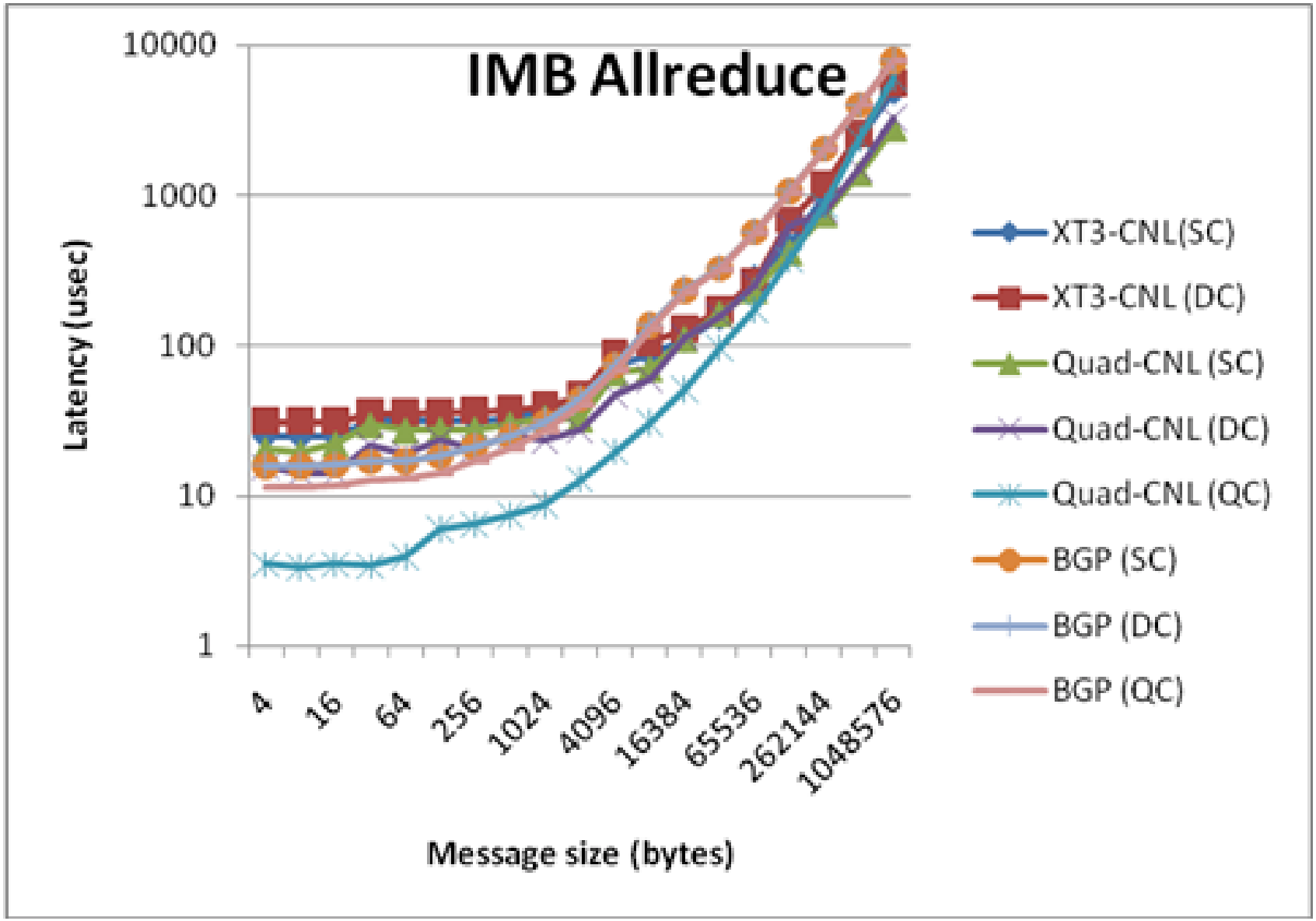
# Performance Challenges

- Is it memory contention?
- Network contention?
- Will it help to make multi-core aware?
- Won't the vendor's "just fix it"?
- What of the HPCS languages?

# Performance Challenges

- Is it memory contention?
- Network contention?
- Will it help to make it multi-core aware?
- Won't the vendor's "just fix it"?
- What of the HPCS languages?

# Performance Challenges



# Performance Challenges

- Is it memory contention?
- Network contention?
- Will it help to make multi-core aware?
- Won't the vendors "just fix it"?
- What of the HPCS languages?

# Performance Challenges

- Is it memory contention?
- Network contention?
- Will it help to make multi-core aware?
- Won't the vendor's "just fix it
- What of the HPCS languages?

## Wednesday

**14 Technical Sessions**

| | 14A Performance | 14B Applications | |
|---|---|---|---|
| 3:15 | **Application Performance on the UK's New HECToR Service**, Fiona Reid, HPCX Consortium (HPCX) | **Exploring the Performance Potential of Chapel in Scientific Computations**, Richard Barrett, Oak Ridge National Laboratory (ORNL) | **C In** |
| 3:45 | **Investigating the Performance of Parallel Eigensolvers on High-end** | **Massively Parallel Electronic Structure Calculations with Python Software**, Jussi | **Fi (H** |

# Summary

- Jaguar has (again) been upgraded:
  - QuadCore processors, SeaStar2, DDR-2
  - 7,832 QC 2.1 GHz procs, 263 TFLOPS : 4 ops/clock!
  - 33.6 GFLOPS / QC

# Summary

## Quad-core faster than dual core

## Performance is not free.

- -tp barcelona-64 : Use it!
- SSE4 crucial.
- mpt3 significant improvement.
- Explore alternative algorithms.
- Try different compilers, options.
- Keep profiling code!

# Acknowledgments

This research was sponsored by the Office of Mathematical, Information, and Computational Sciences, Office of Science, U.S. Department of Energy under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

Our work reported in this talk and paper was greatly aided by:

Cray, including the Center of Excellence at ORNL:
Jeff Beckleheimer, Jeff Larkin, John Levesque, Cathy Willis

Brian Waldecker of AMD, Inc.

Apps:
AORSA: Fred Jaeger  Gyro: Jeff Candy
S3D: Jackie Chen

ORNL
Sean Ahern, Ed D'Azevedo, Don Maxwell, Arnold Tharrington; all members of Scientific Computing group and many others in NCCS.