Scaling Properties of the M3D Code From CDX to ITER

Jin Chen
M3D Group

FDM3D Workshop, Princeton

March 19, 2007

Background: Large scale sparse linear systems

[1] $1N: \ \Delta^{\perp} u = f^{\perp}$, Neumann boundary condition

[2] $1D: \ \Delta^{\perp} u = f^{\perp}$, Dirichlet boundary condition

[3] $2D: \ \Delta^{*} u = f^{*}$, Dirichlet boundary condition

[4] $3D: \ \Delta^{\dagger} u = f^{\dagger}$, Dirichlet boundary condition

[5] $4N: \left(\Delta^{\perp} - \eta\right) u = f_g^{\perp}$, Neumann boundary condition

[6] $4D: \left(\Delta^{\perp} - \eta\right) u = f_g^{\perp}$, Dirichlet boundary condition

[7] $4I: \left(\Delta^{\perp} - \eta\right) u = f_g^{\perp}$, semi-implicit advancing of thermal conduction

[8] $5D: \left(\Delta^{*} - \eta\right) u = f_g^{*}$, Dirichlet boundary condition

[9] $6D: \left(\Delta^{\dagger} - \eta\right) u = f_g^{\dagger}$, Dirichlet boundary condition

Zero eigenvalue, Ill conditioned,
Null-space,
CG/ICC, Slow convegence, CG/AMG

$$\Delta^{\perp} \equiv \frac{\partial^2}{\partial R^2} + \frac{\partial^2}{\partial Z^2}$$

$$\Delta^{*} \equiv \Delta^{\perp} - \frac{1}{R}\frac{\partial}{\partial R} \quad \textit{unsymmatric}$$

$$\Delta^{\dagger} \equiv \Delta^{\perp} + \frac{1}{R}\frac{\partial}{\partial R}$$

$$\Downarrow$$

Conserved form

$$\Delta^{\perp} \equiv \nabla \cdot \nabla$$

$$\Delta^{*} \equiv \nabla \cdot \frac{1}{R} \nabla \quad \textit{symmetric}$$

$$\Delta^{\dagger} \equiv \nabla \cdot R \nabla$$

Majority of the code time spent in these Linear Elliptic solvers:

~80%, GMRES/ILU

~33%, CG/AMG

significant speedup was observed for large problems, very scalabe.

Motivation: WHY

To optimize the code and prepare for petascale calculations.

M3D time can be broken down to 3 major parts:

scales up to thousands of processors

	characteristics		
Physics	Do loops		
Linear Solver	matrix invert	**gmres/ilu** →	**cg/hypre**
Data copy	cross processor communications		

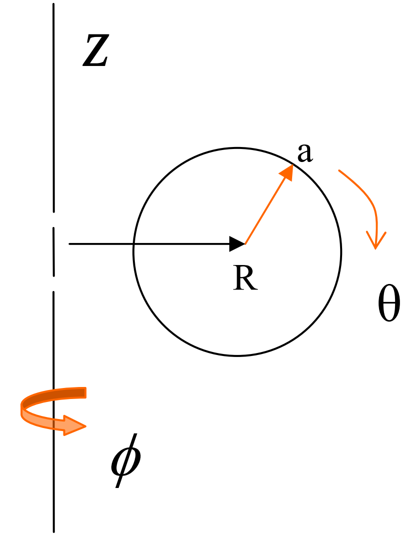
Their efficiencies are critical for optimization on petascale computers.

An Example: Total M3D Time = 726 sec

Physics = 240
Linear Solver = 274 (optimized, otherwise >80%.)
Data copy = 206

Outline: HOW

1. 3D (r,θ,φ) *strong* scaling
2. 3D (r,θ,φ) *weak* scaling
3. 1D (φ) *weak* scaling
4. 2D (r,θ) *weak* scaling

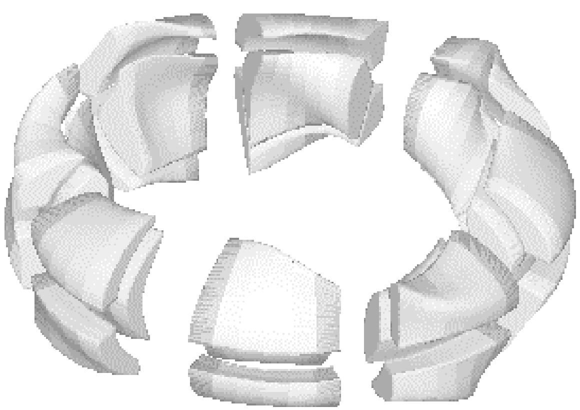


A, B, C, D, E, F

A geomotry in in toroidal φ direction;
B CPUs in toroidal φ direction.

C geometry in minor radial r direction;
D CPUs in radial r direction.

E geometrt in poloidal θ direction;
F CPUs in poloidal θ direction.



Definitions

Physics grid: A C E

CPU grid: B D F

1D (φ) *weak* scaling

A	B	C	D	E	F
a1	b1				
a2	b2				
:	:				
an	bn				

2D (r,θ) *weak* scaling

A	B	C	D	E	F
		c1	d1	e1	f1
		c2	d2	e2	f2
		:	:	:	:
		cn	dn	en	fn

3D (r,θ,φ) *weak* scaling

A	B	C	D	E	F
a1	b1	c1	d1	e1	f1
a2	b2	c2	d2	e2	f2
:	:	:	:	:	:
an	bn	cn	dn	en	fn

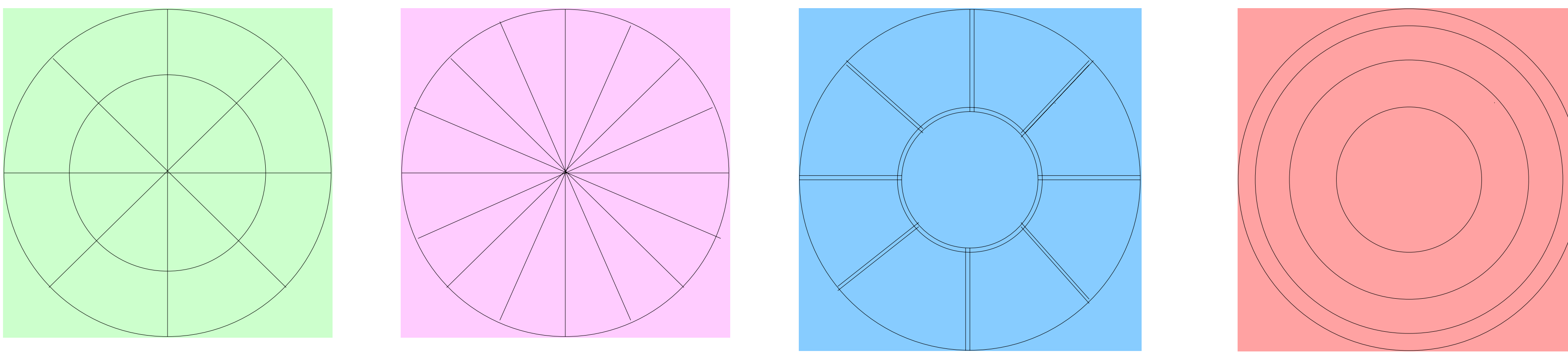
3D (r,θ,φ) *strong* scaling

A	B	C	D	E	F
	b1		d1		f1
	b2		d2		f2
:	:	:	:	:	:
	bn		dn		fn

Strategy to improve data copy

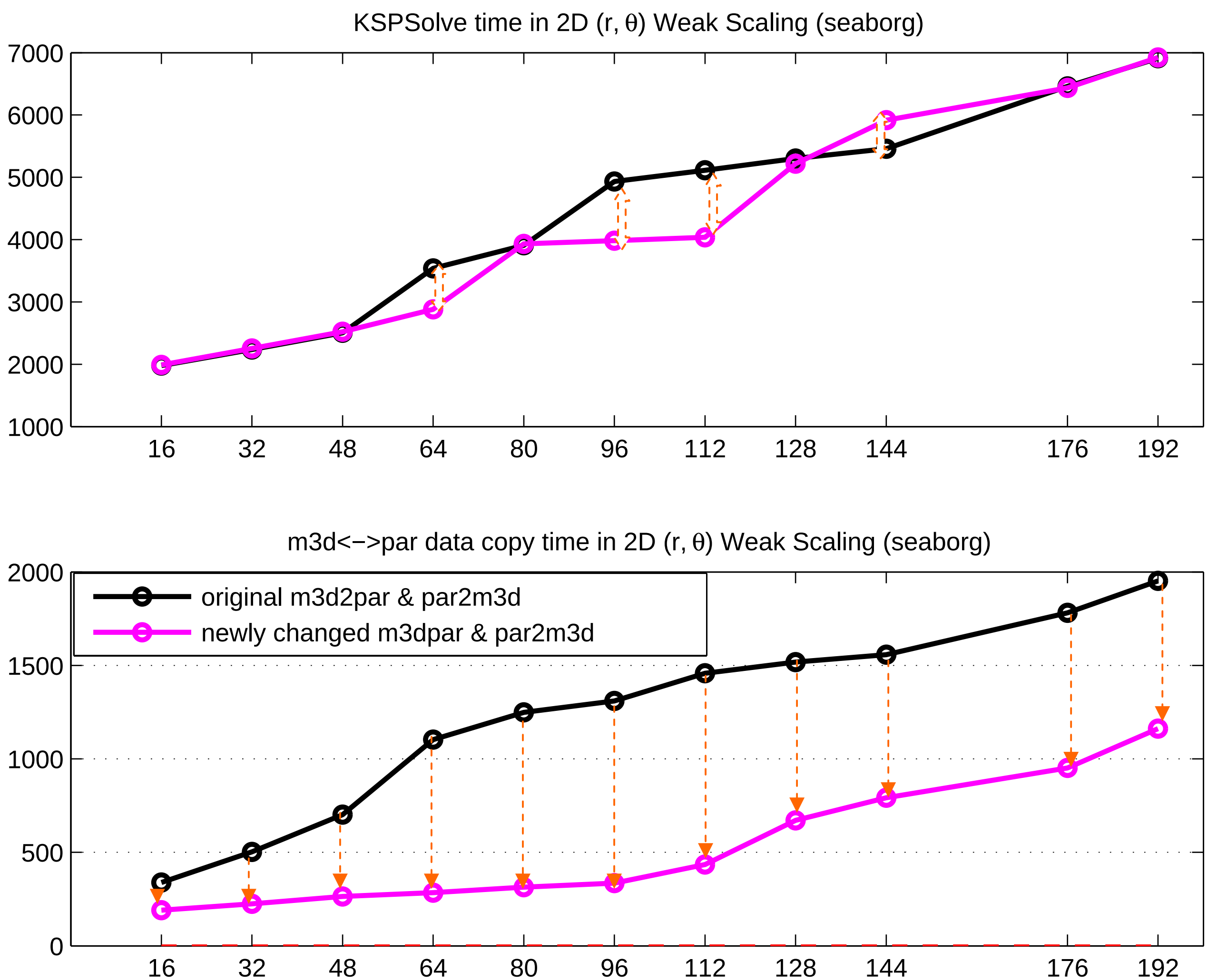
a) Reduce toroidal ghost changes
b) Reduce poloidal ghost changes:

2 times faster on seaborg



2D (r,θ) direction, C D E F

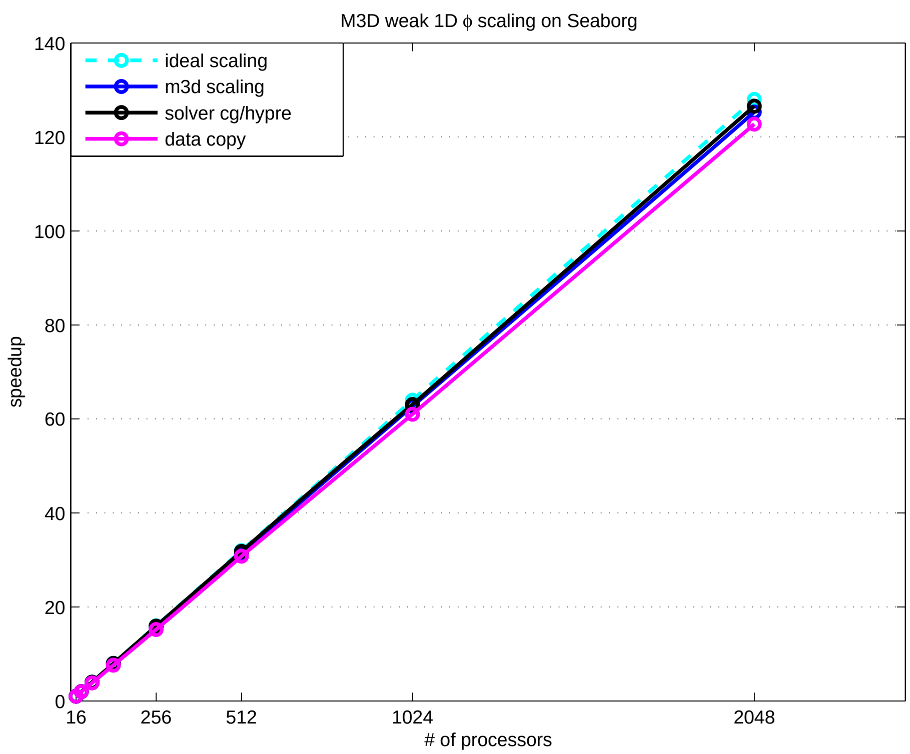
Strategy to improve data copy – II

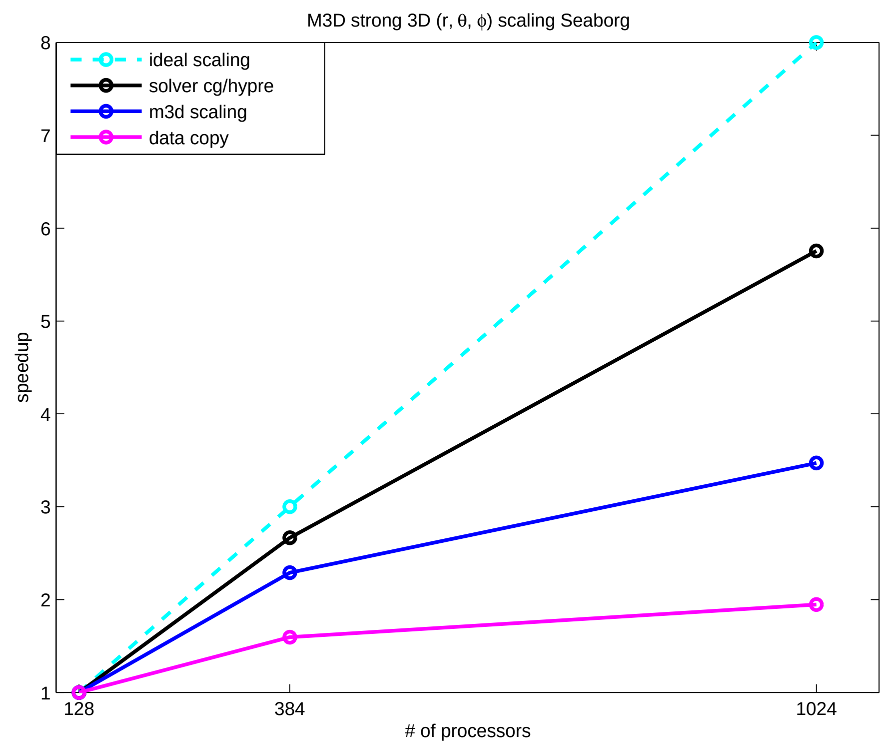


Seaborg: NERSC IBM SP RS/6000.

a distributed memory computer with 6,080 processors.
Each processor has a peak performance of 1.5 GFlops.
The processors are distributed among 380 compute *nodes*
with 16 processors per node. Processors on each node
have a shared memory pool of between 16 and 64 GBytes

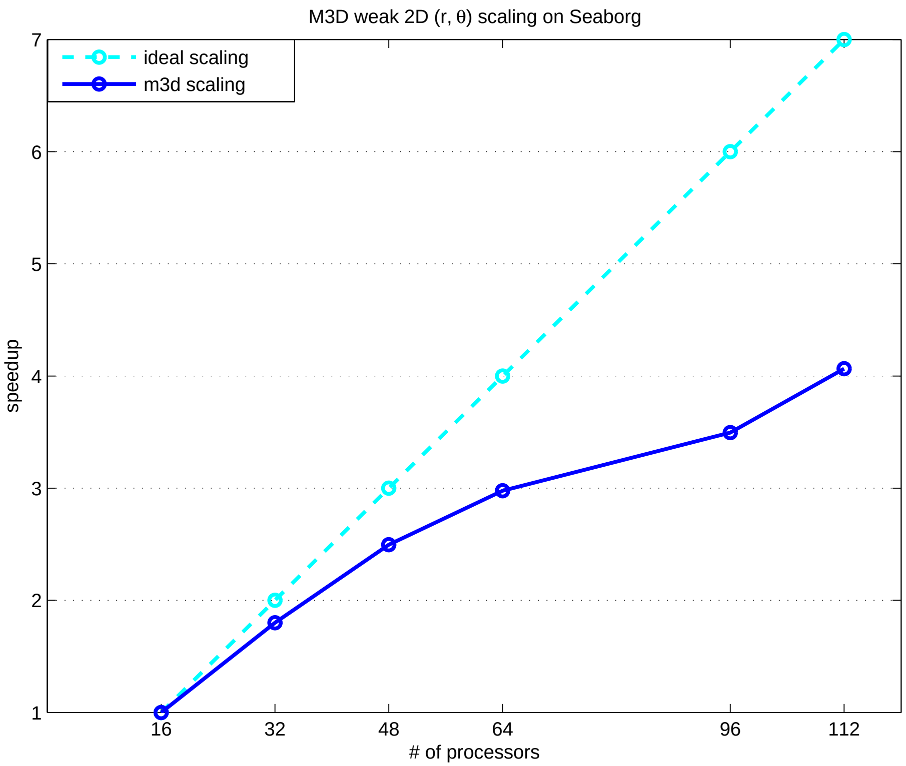
1D weak & 3D (r,θ,φ) strong scaling-till 10/29/2006

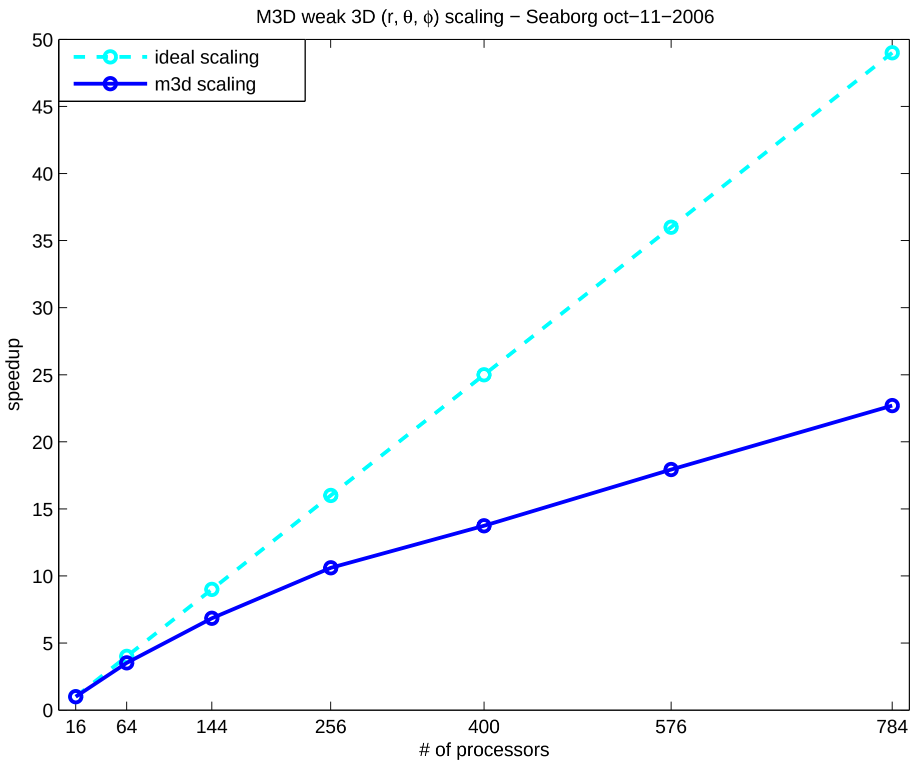




1D good; 3D not as good as 1D and can only go up to 1024 procs.

2D & 3D (r,θ,φ) weak scaling



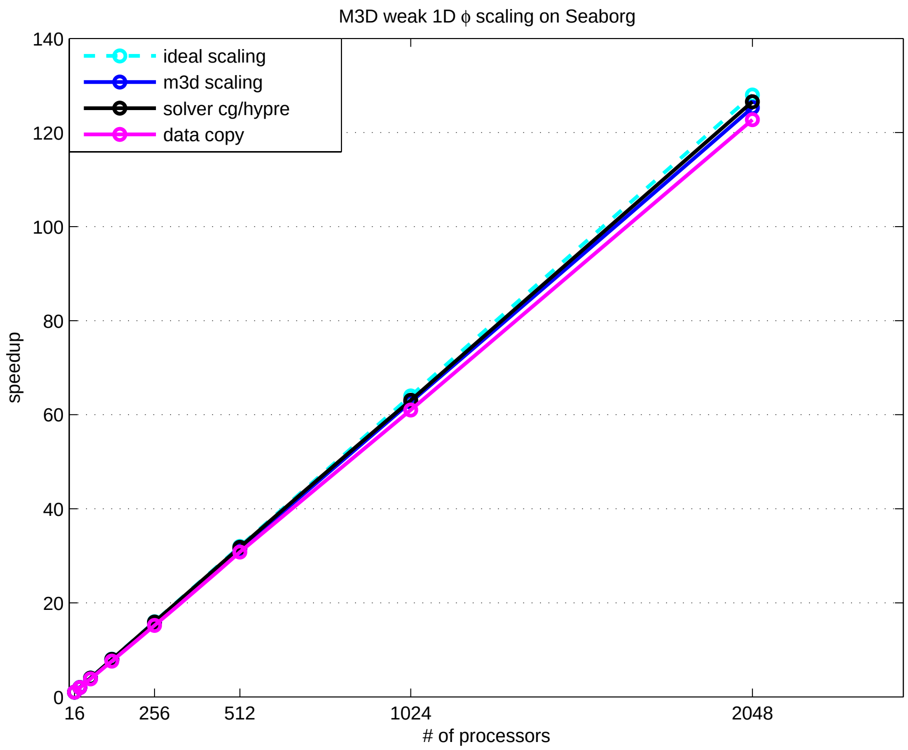


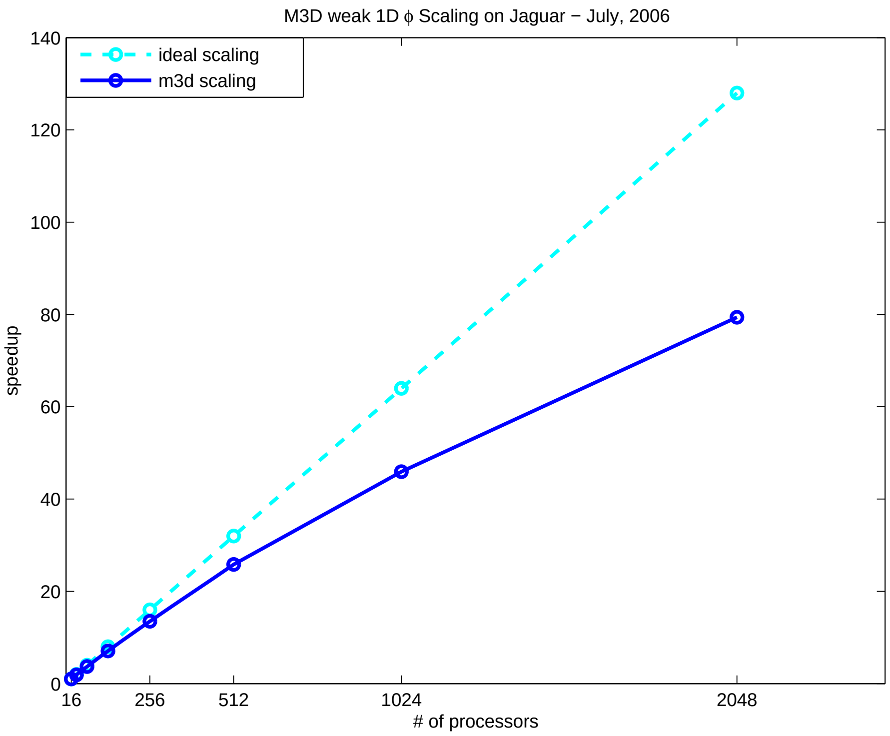
2D and 3D not as good as 1D, and crashes beyond 112 or 784 procs.

Jaguar: XT3

Compute-node processor count	10,424 cores Note: *2 CPUs now share the memory and interconnect bandwidth of a single CPU before the upgrade*
Compute-node processor size	2.6 GHz dual core
Compute-node memory	4 GB *Maintaining 2 GB per core*
Lustre file system capacity	100 TB
Luster default stripe width	4 OSTs *The stripe size can be changed with the lfs stripesize command*
UNICOS/lc	Upgraded to 1.4.22 *Executables must be recompiled*
Interconnect	Full 3D torus

1D (φ) weak scaling-till 7/15/2006





seaborg

Jaguar: not comparable to seaborg

Problems fixed on Jaguar

- Runtime memory limitation
  - Solution: use only 1 processor per node

yod –SN m3dp_fsymm_opt.x …

- Code crashes when the number of processor increases from 2048 to 3076 or 4096:

module load gmalloc

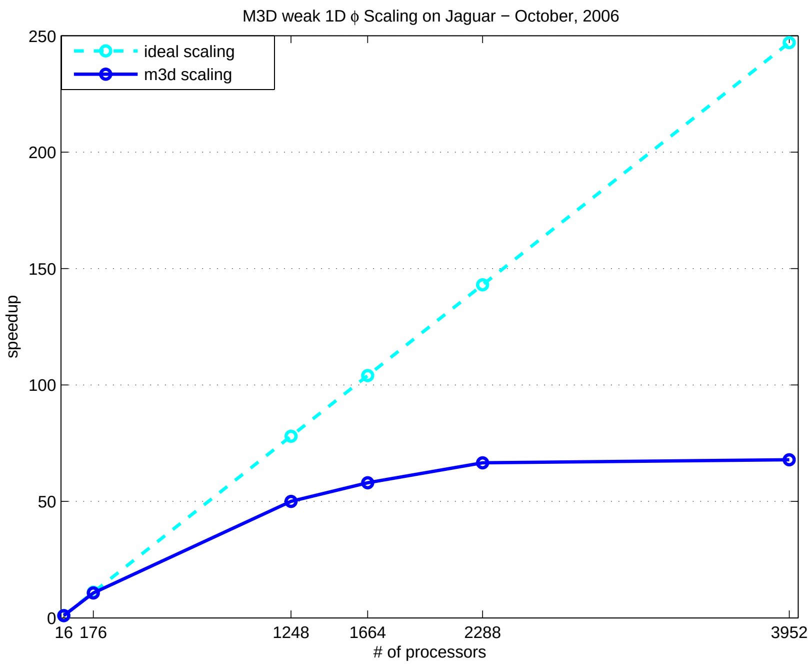
link –gmalloc as the last library to build m3dp.x

- Wait too long when debugging code
  - We need dedicated time to fix bugs only appeared on large number of processors.
- Fortran static array (stack)

yod –SN –stack 500M m3dp_fsymm_opt.x …

All the problems were fixed after 9/15/06 upgrades.

1D (φ) weak scaling – till 10/26/2006



Finally we can go up to 4000 procs by the end of October, 2006.

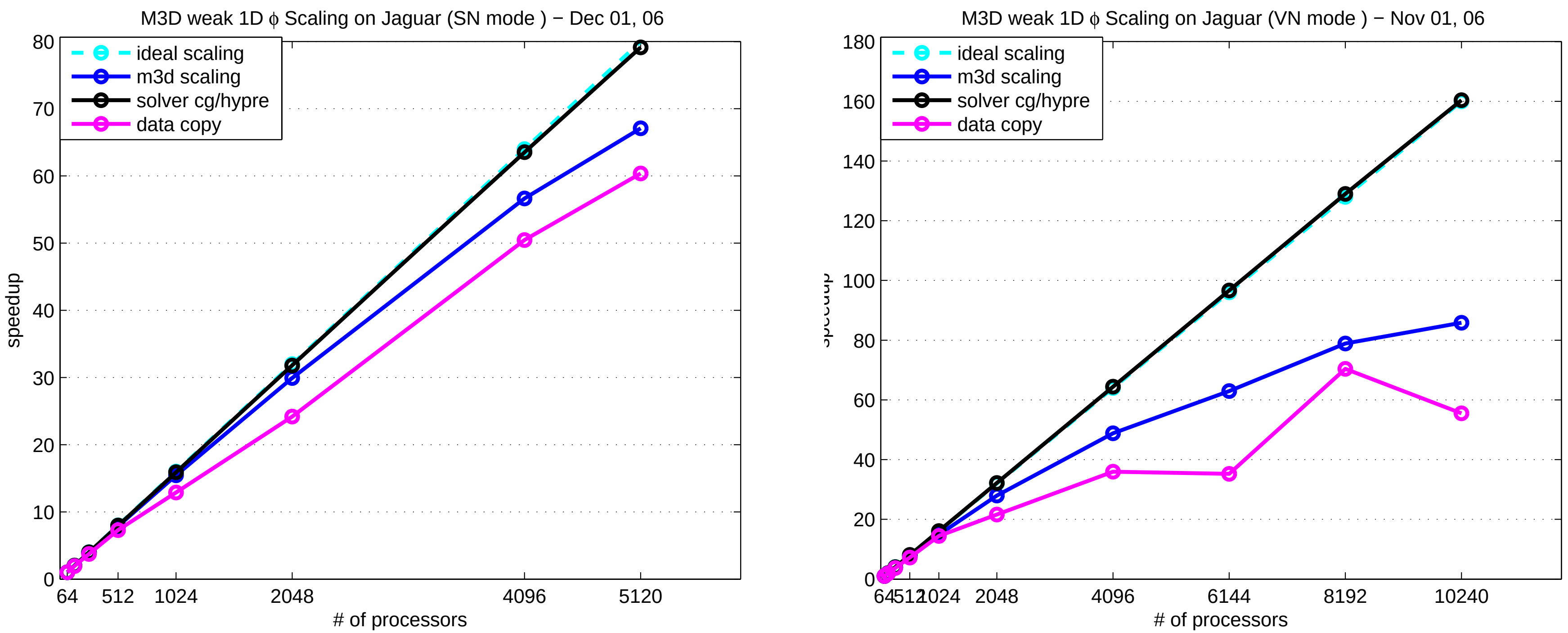
Communication:

A	B	C	D	E	F
x	x	51	1	4	4
16					
176					
1248					
1664					
2288					
3952					

Heavy communication in φ direction compared to (r,θ) directions.

MPI I/O

1D weak scale on the whole machine

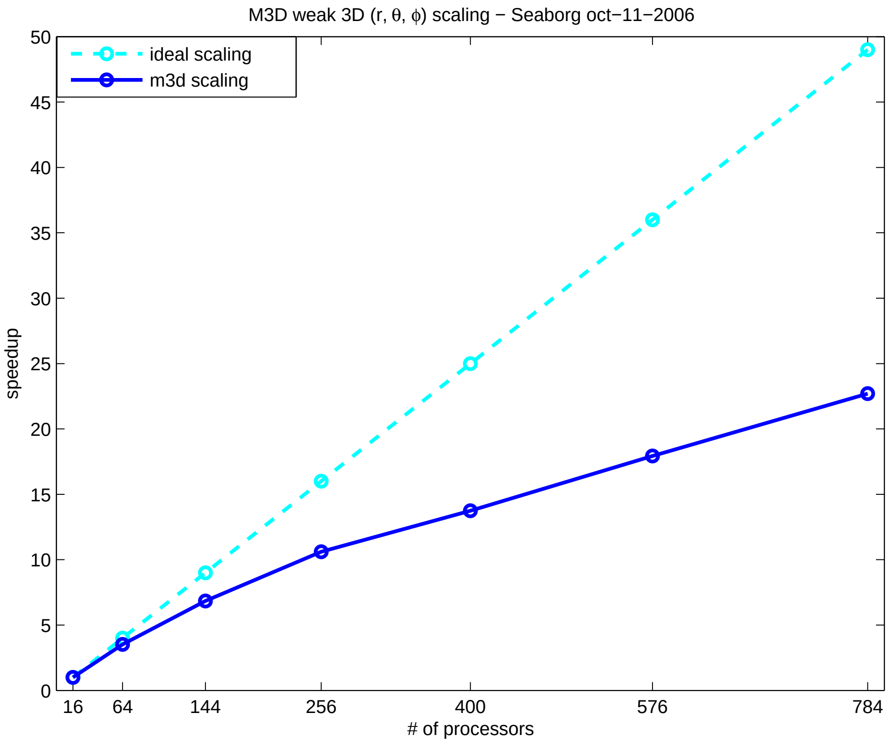


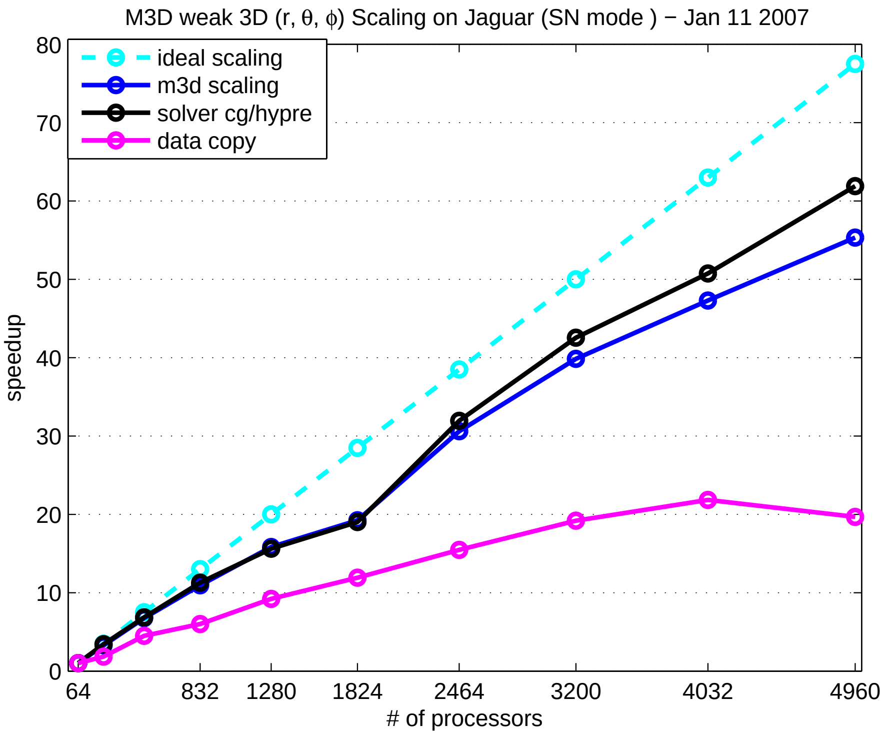
Good up to 5000 processors; runs on > 10,000 proces, although not ideal.

Problem size:

Vert=C(C-1)/2*E+1=62,6081/plane, total=(16, 32, 64, 128, 256, 512, 1024, 2048) times of Vert.

3D weak scale on the whole machine

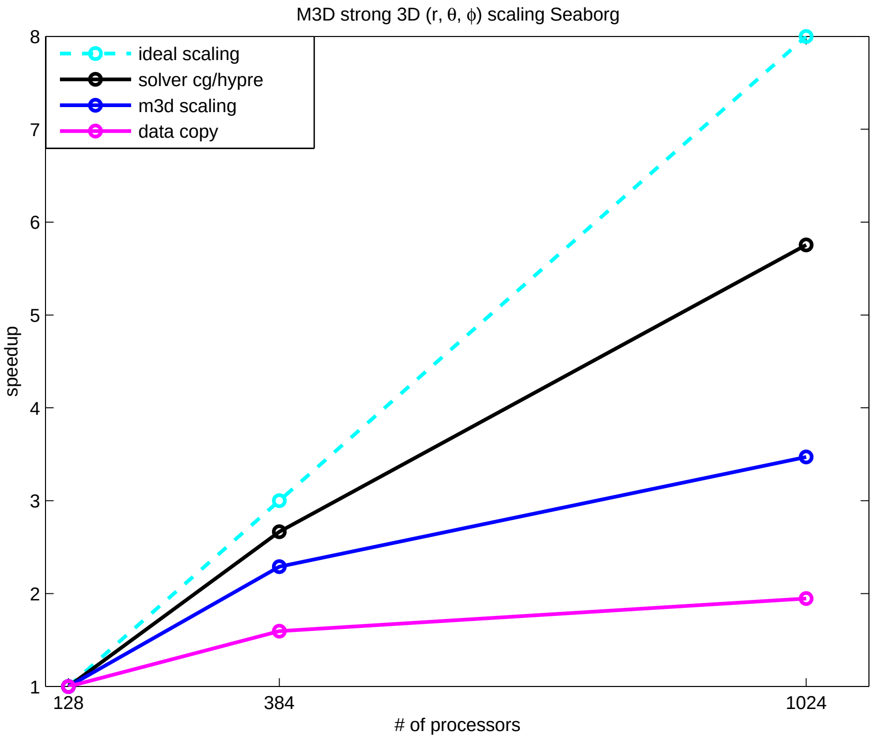




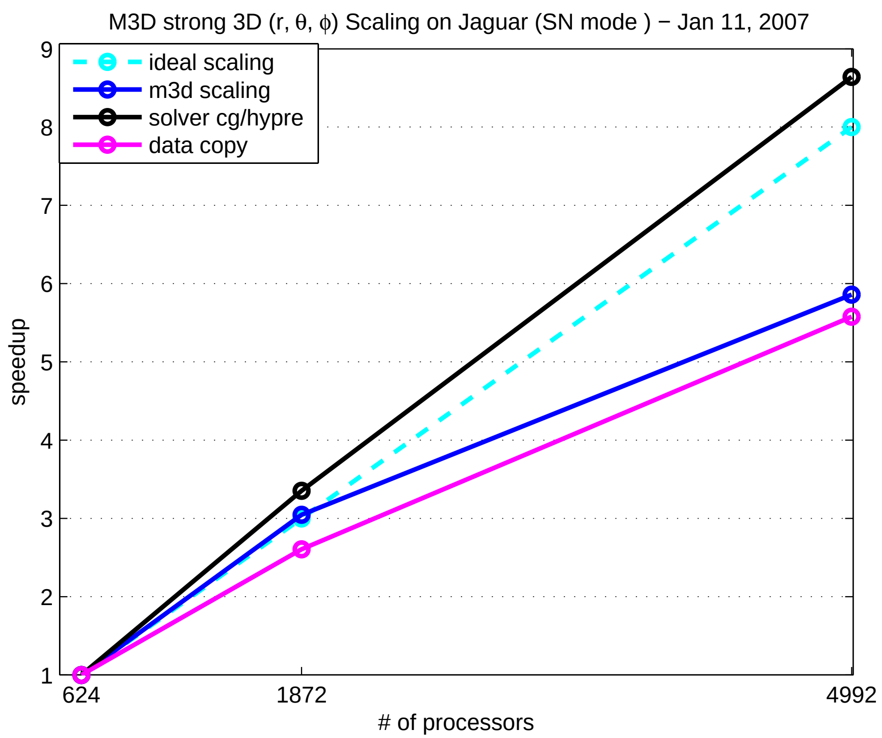
Seaborg: ~800 procs

Jaguar: ~5000 procs

3D strong scale on the whole machine



Seaborg: 1024 procs



Jaguar: 5000 procs

BGL at Argonne

MCS BGL Configuration

Compute - 1024 dual PowerPC 440 700MHz 512MB nodes

Storage - 14 TB of clusterwide disk (currently using the MCS Parallel Virtual File System (PVFS)) and 3.5TB of home directory filespace.

Network - IBM BlueGene Torus, Global Tree and Global Interrupt

Running Jobs

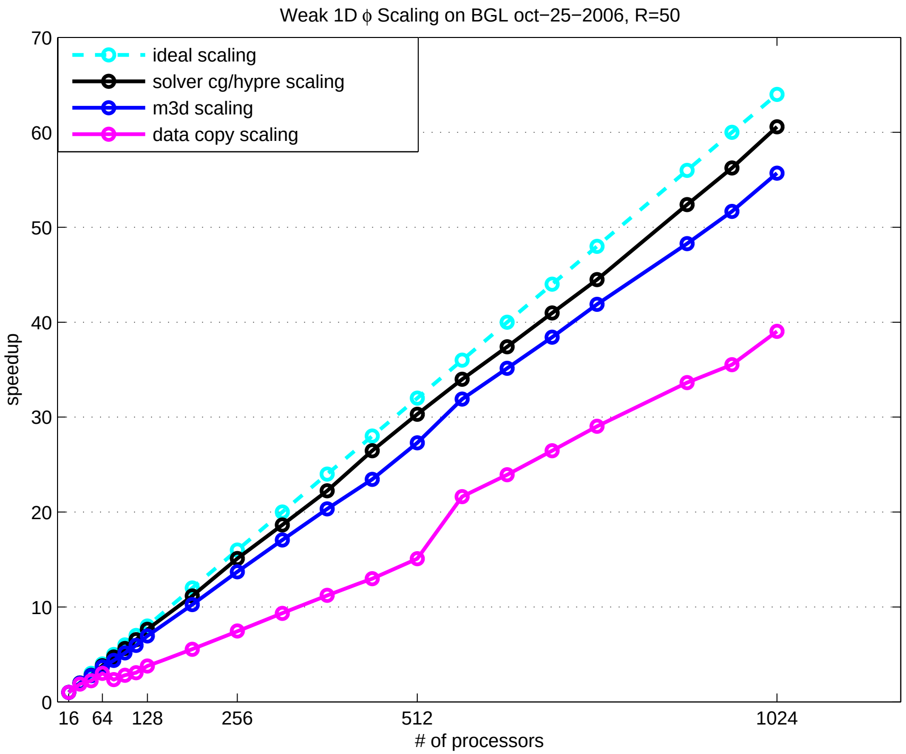
cqsub -t <time> -n <nodecount> -c <#processors> -m <mode> <exe> [arg1,arg2,...]

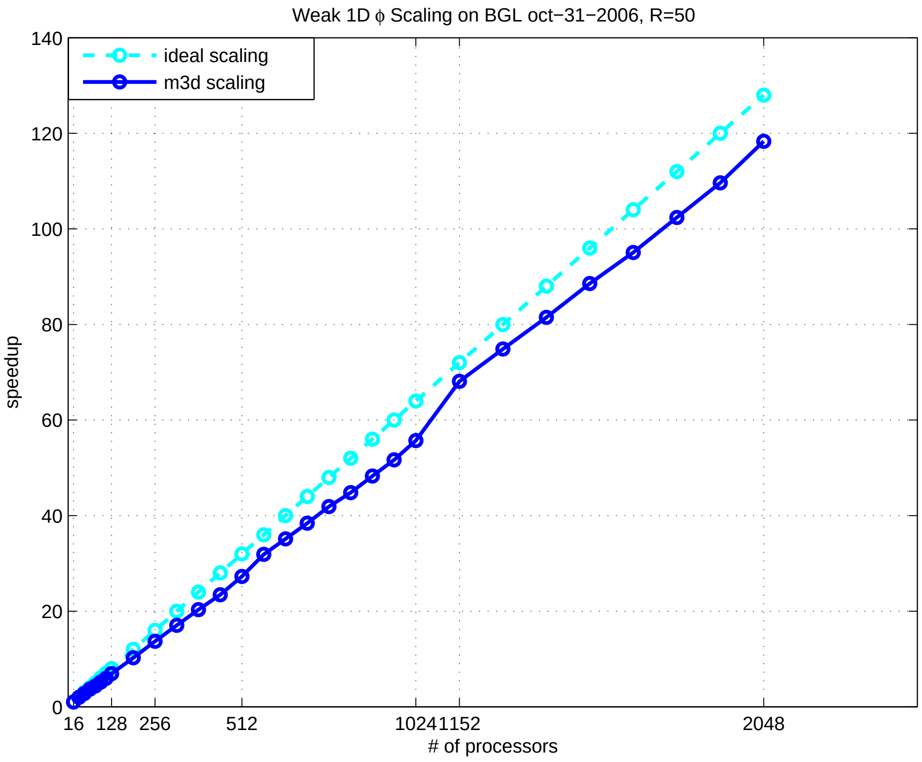
<time> is in minutes (required)
<nodecount> is the number of nodes
<#processors> number of procs
<mode> is one of 'co' or 'vn'
<exe> is the full path name to the mpi executable **[arg1,arg2,...]**

using a partition size smaller than 512, the code cannot use the full Torus network. This will most likely cause the performance to be very poor.

Desired Usage	Partition Size	# of processors
Development, Scaling	32	64
Development, Scaling	64	128
Development, Scaling	128	256
Development, Scaling	256	512
No Development	512	1024
No Development	1024	2048

1D (φ) weak scaling





Doesn't have enough memory to push big runs.

As a result of scaling, code stretching …

Meaningful,
Not nonsense!

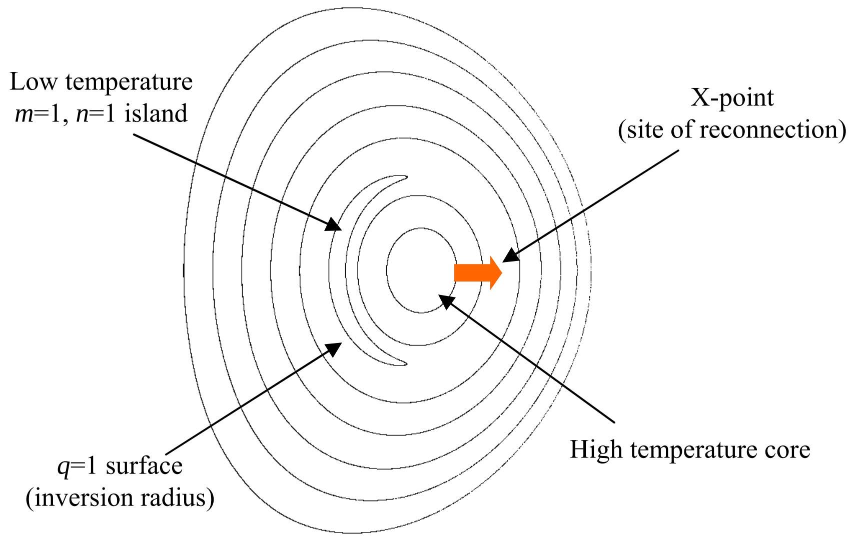
- Fixed
  - physics bugs
  - memory bugs (not show up when using a small amount of memory)
  - coding bugs (not show up when Vert is small)
  - I/O bugs

- Robust (ready to do comprehensive physics and get higher resolution)

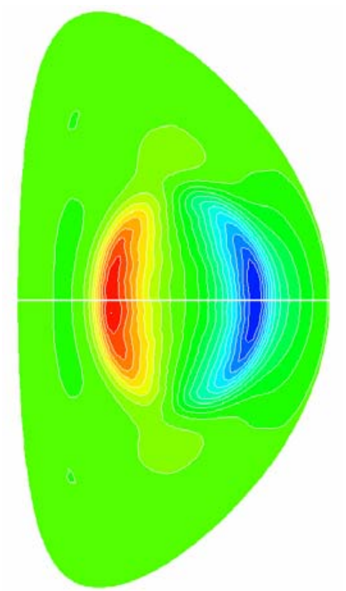
(A, B, C, D, E, F) = (32, 16, 51~91, 1, 4, 4), Vert= 5,100/(r,θ) plane
↓
(A, B, C, D, E, F) = (3952, 640, 560, 2, 13, 39), Vert=2,034,760/(r,θ) plane

Franklin : Applications need high resolution

CDXU --> DIIID --> ITER



Sawtooth Instability in CDXU



Nonlinear behavior of energetic particle modes in NSTX