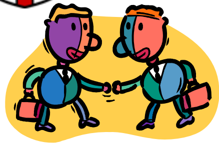
A Multilayer Approach to Simulate Large Multiscale Computational mechanics Problem Using Grid



Brian Toonen
Nicholas Karonis

Leopold Grinberg
George Em Karniadakis

Joseph Insley
Michael Papka

This work was supported by the National Science Foundation's CI-TEAM program under contract OCI-0636412, NSF NMI program under contract OCI-0330664 and NSF-NMI 0330664.

Also NSF programs directed by Leland Jameson, Frederica Darema, and Almadena Chtchelkanova.

The 20+ Year Vision

- Imagine a "digital body double"
  - 3D image-based medical record
  - Includes diagnostic, pathologic, and other information
- Used for:
  - Diagnosis
  - Less invasive surgery-by-robot
  - Experimental treatments

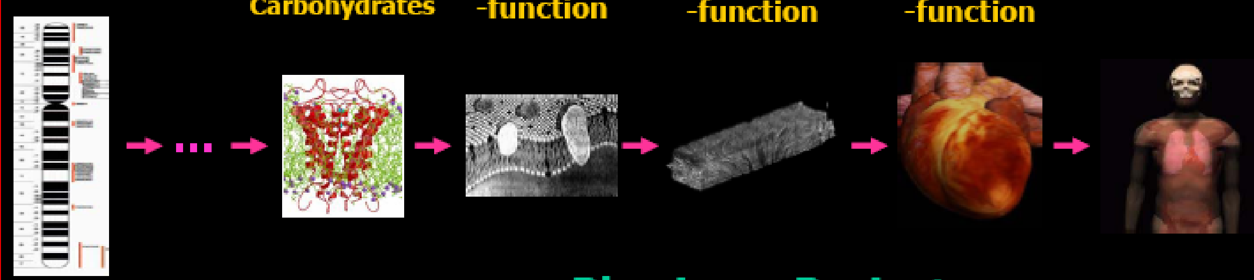
- Digital Human Effort
  - Lead by the Federation of American Scientists

Digital Human: International Project

http://www.fas.org/dh/

Genes to Organs & Organisms

Genes | mRNA | Proteins Lipids Carbohydrates | Cell structure -function | Tissue structure -function | Organ structure -function | Clinical medicine



Physiome Project

Genome/Transcriptome/Metabolome/Proteome

20,000 genes → 60,000+ proteins → 200+ cell types → 4 tissue types → 12 organ systems → 1 body

Hunter, PJ and Borg, TK. Integration from proteins to organs: The Physiome Project. Nature Reviews Molec& Cell Biol.4:237-243, 2003

How Large is the Arterial-Tree Problem?

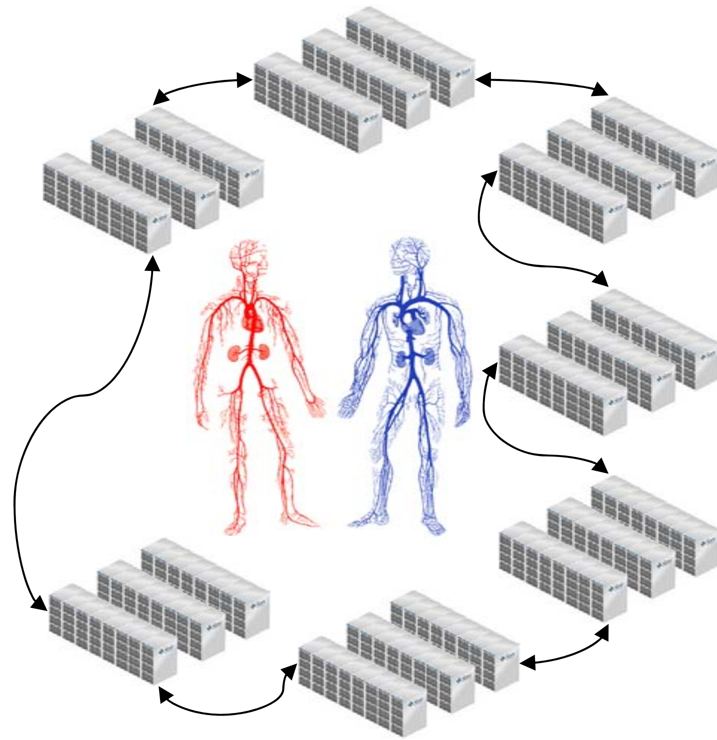
- On the average, an adult human who weighs 70 kg. has a blood volume of 5 liters.

Typical volume of tetrahedral elements with an edge of 0.5 mm is about 0.0147 mm^3.

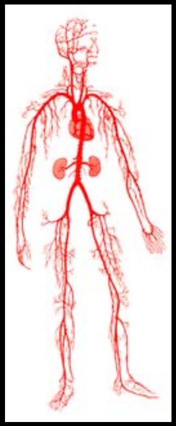
- 339M tetrahedral elements.

- 339E6*(P+3)(P+2)^2 = **85.4 E9** Degrees of Freedom per one variable (P=4) => **10 TB of Memory!**

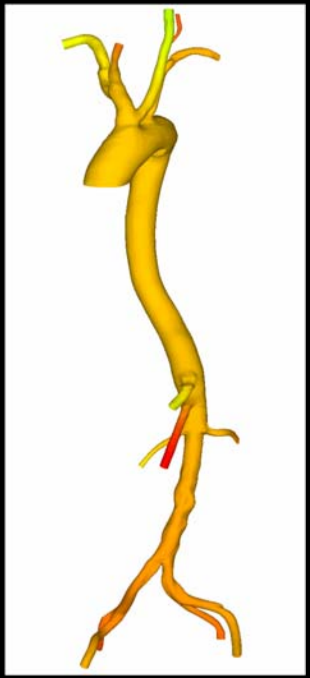
- ***A human is a multicellular eukaryote consisting of an estimated 100 trillion cells…***



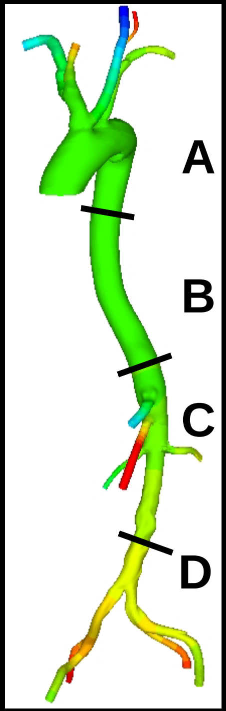
3D Numerical Simulation of Flow in Human Aorta



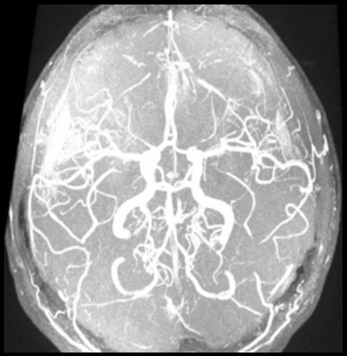
	# of spectral elements	Polynomial order	# of DOF (per variable)
Domain A	120,813	6	69,588,288
Domain B	20,797	6	11,979,072
Domain C	106,219	6	61,182,144
Domain D	77,966	6	44,908,416
Total	**325,795**	**6**	**187,657,920**



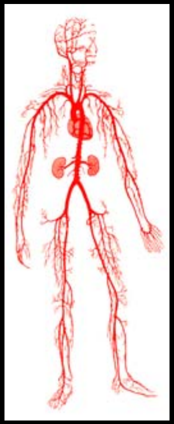
# CPU (CRAY XT3)	256	994	1976
CPU time / time step	4.06 sec	1.24 sec	0.77 sec

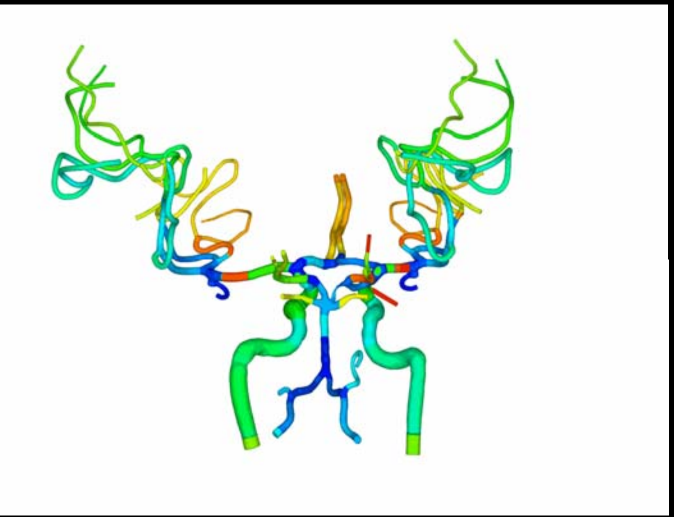


3D Numerical Simulation of Flow in Human Cranial Arterial System

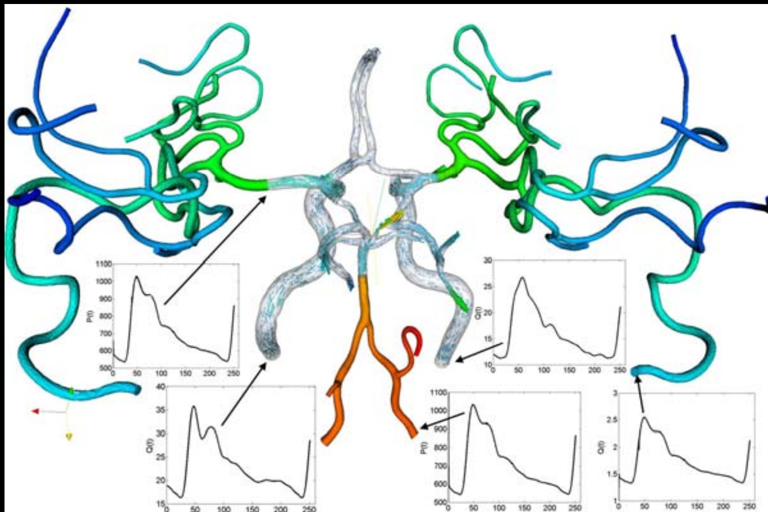


	# of elements	Polynomial order	# of DOF per variable
Domain A	162,909	5	63,860,328
Domain B	44,632	5	17,495,744
Domain C	128,508	5	50,375,136
Domain D	123,201	5	48,294,792
Total	**459,250**	**5(6)**	**180,026,000 (264,528,000)**

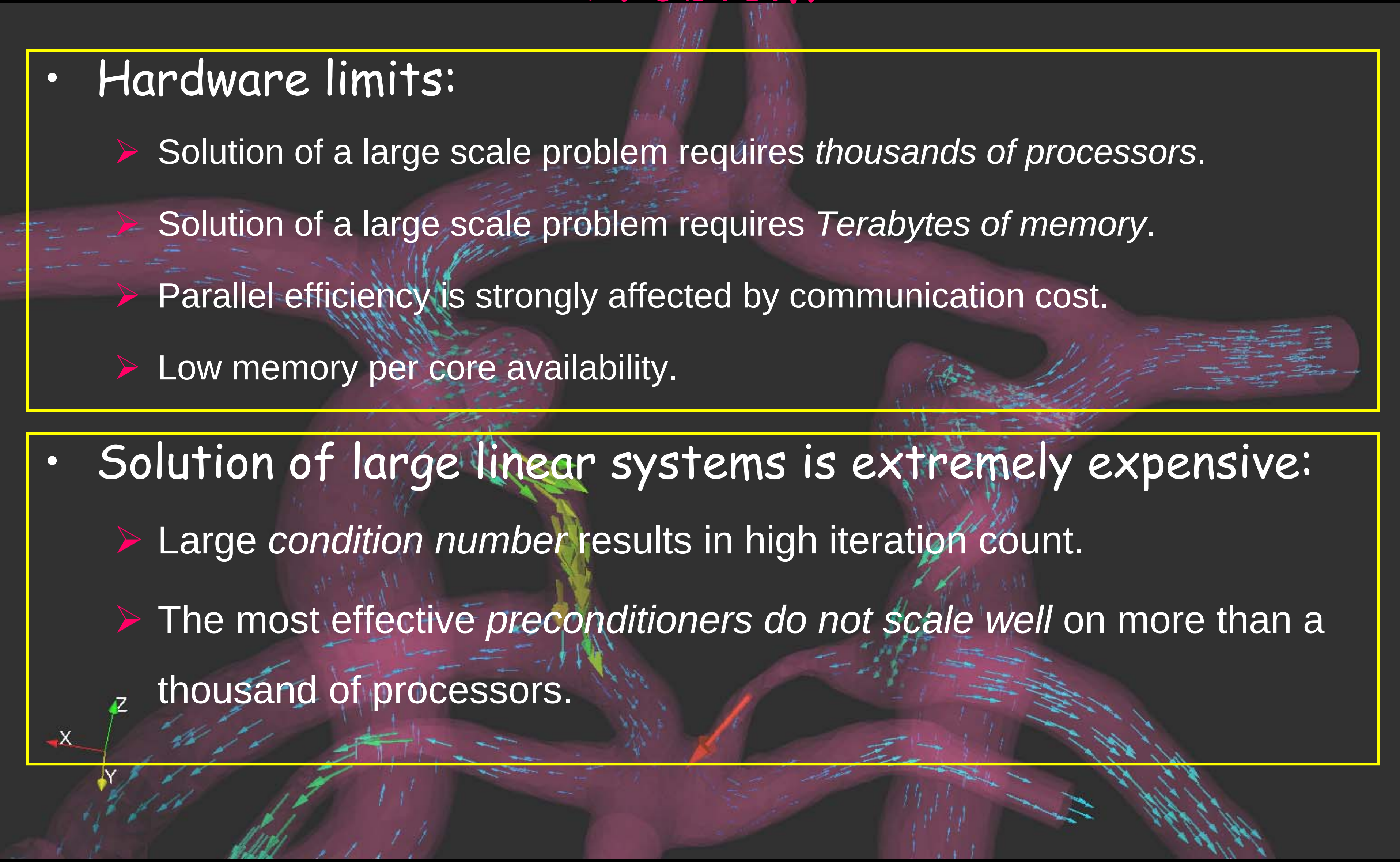




# CPU (CRAY XT3)	CPU-time / time step
1024	1.4 sec
2048	0.92 sec
3265	0.61 sec



Impediments to Solution of Large Scale Problem



- **Hardware limits:**
  - Solution of a large scale problem requires *thousands of processors*.
  - Solution of a large scale problem requires *Terabytes of memory*.
  - Parallel efficiency is strongly affected by communication cost.
  - Low memory per core availability.

- **Solution of large linear systems is extremely expensive:**
  - Large *condition number* results in high iteration count.
  - The most effective *preconditioners do not scale well* on more than a thousand of processors.

<u>Our solution</u>: a multi-layer hierarchical approach.

- Software: NEKTAR-G2
- Multilevel Partitioning
- Solution of Large Scale Problem:
  - on a single supercomputer
  - on TeraGrid

NEKTAR-G2

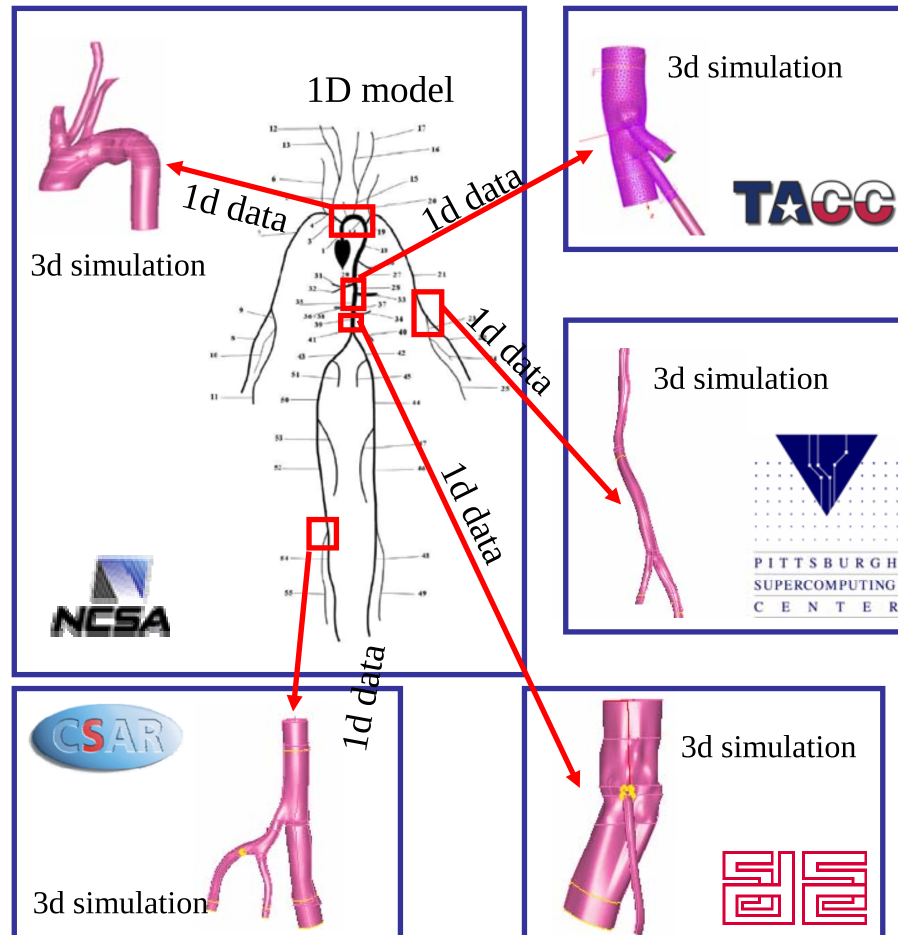
- NEKTAR-G2 Prototype
- The New Domain Decomposition Technique: The Idea

NEKTAR-G2: Prototype[1]

- Overall simulation consists of
  - 1D computation through the full arterial tree;
  - Detailed 3D simulations on arterial bifurcations.

- 1D results feed 3D simulations, providing flow rate and pressure for boundary conditions.

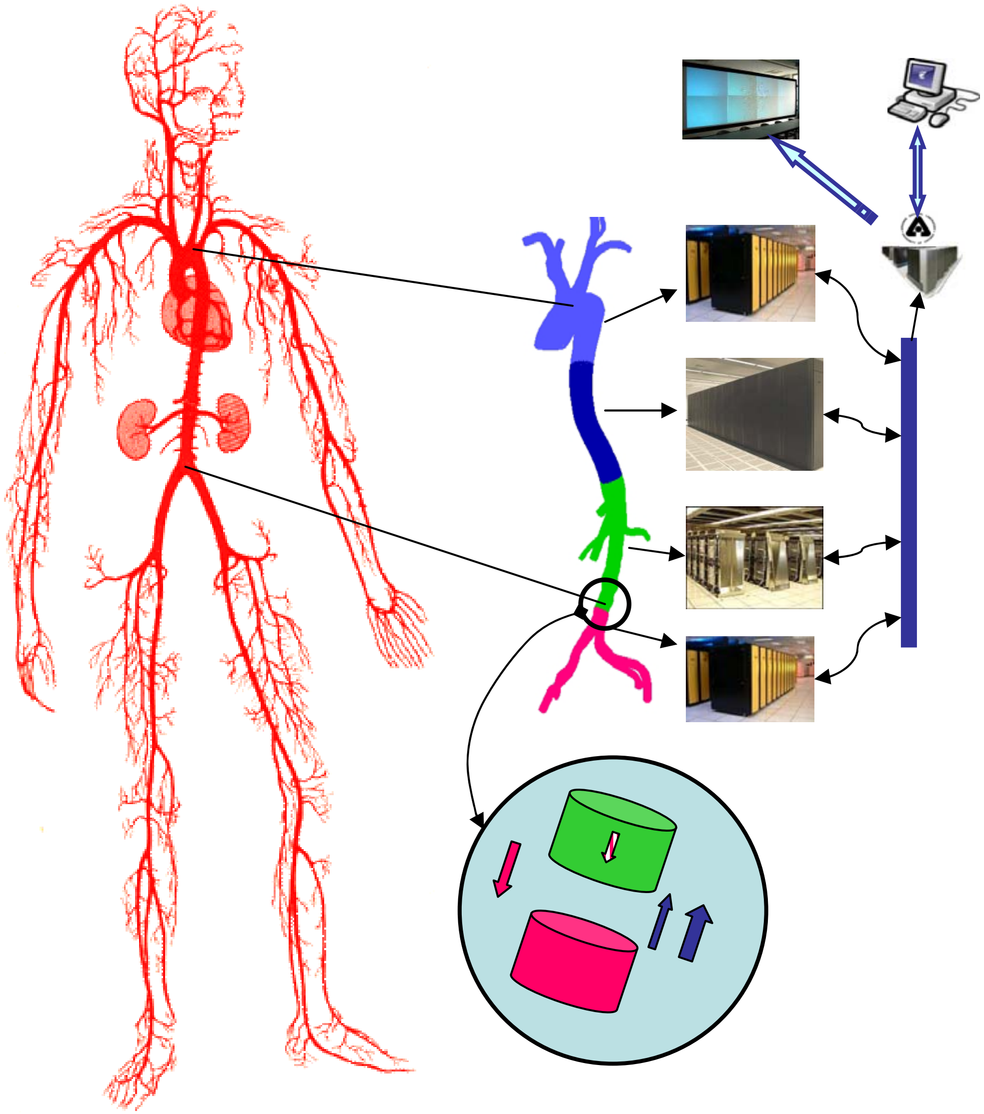
- MPICH-G2[2] was used for intra-site and inter-site communications on TeraGrid.



[1] S. Dong *et al.*, "Simulating and visualizing the human arterial system on the TeraGrid", *Future Generation Computer Systems*, Volume 22, Issue 8, October 2006, pp. 1011 - 1017

[2]N. Karonis et al, A Grid-Enabled Implementation of the Message Passing Interface, (*JPDC*), Vol. 63, No. 5, pp. 551-563, May 2003

NEKTAR-G2: Large Scale Flow Simulations on the TeraGrid

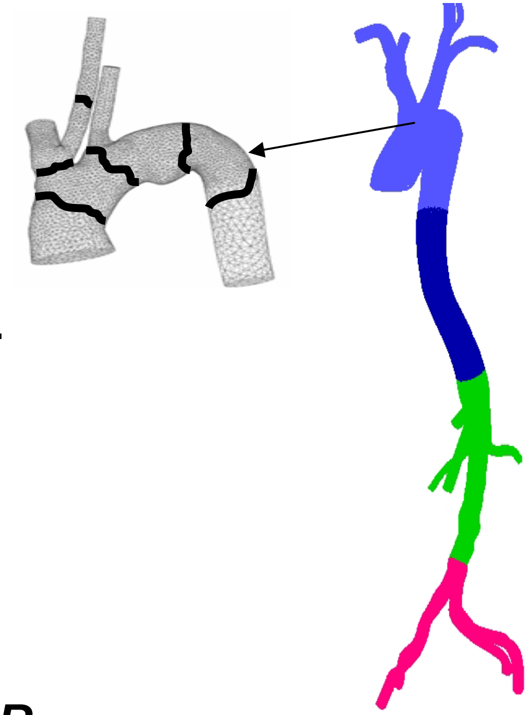


Nektar-G2 features 3D-3D coupling between domains.

Increased volume of data transfer between 3D blocks requires high level of parallelism.

MPIg is used for intra-site and inter-site communications on TeraGrid.

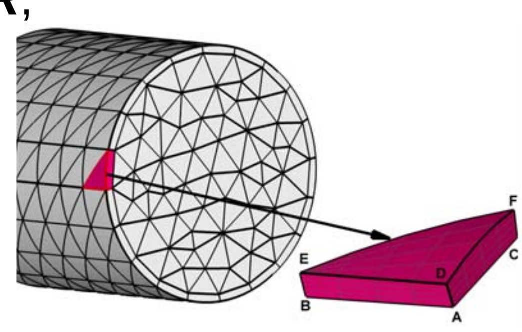
Two-Level Domain Decomposition Technique for TeraGrid Simulations: Method



- Numerical solution is performed on ***two levels***:
  on ***outer level*** loosely coupled problem is solved,
  on ***inner level*** several tightly problems are solved in parallel.

- ***Multi-level partitioning*** of the entire computational domain requires ***multi-level parallelism*** in order to maintain high parallel efficiency using thousands of processors.

- Solution of tightly coupled problems is performed by ***NEKTAR***, a parallel numerical library based on the ***high-order spectral/hp element method***[1].
  Continuity in numerical solution is achieved by imposing proper boundary conditions on the sub-domains interfaces.



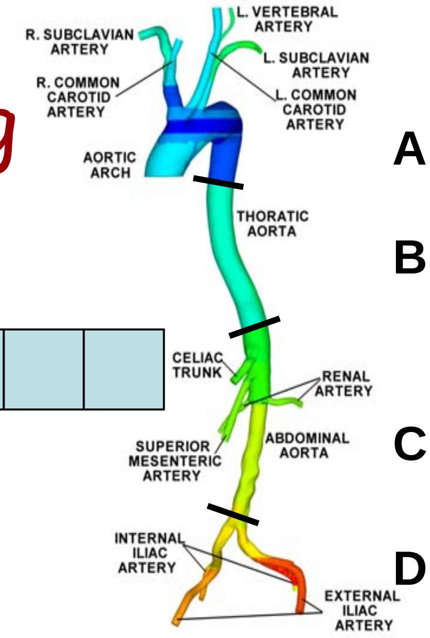
- On TeraGrid ***inter-*** and ***intra-site communication*** is performed by **MPIg library**.

[1] *Karniadakis & Sherwin, Spectral/hp Element Methods for CFD, 2005, 2nd edition, Oxford University Press*

Multilevel Partitioning of Global Communicator

- High Level Communicator Splitting
- Low Level Communicator Splitting
- Message Passing Across Sub-Domain Interface

Two-Level Domain Decomposition:
High Level Communicator Splitting



		M	P	I		C	O	M	M		W	O	R	L	D				

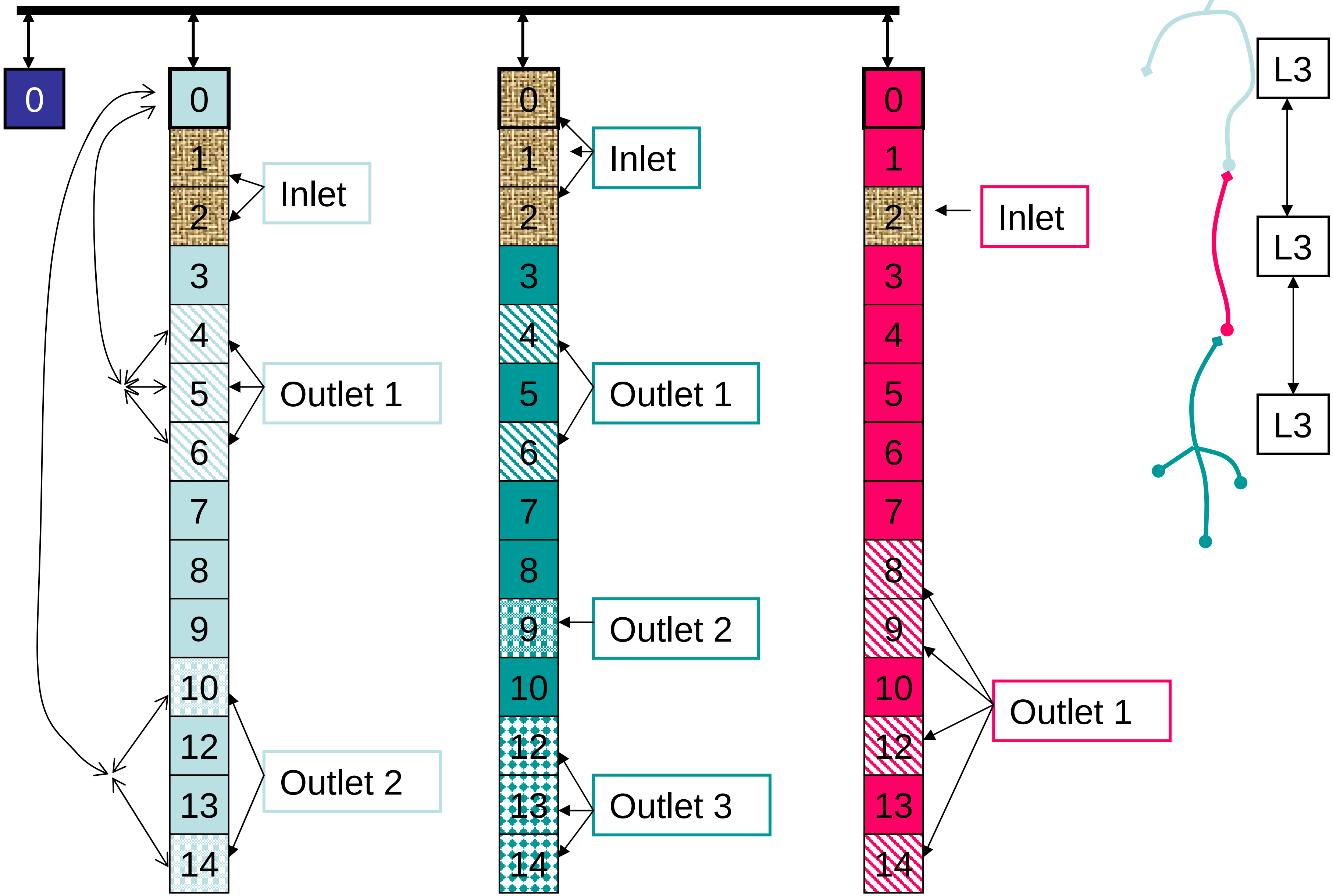
TOPOLOGY AWARE DECOMPOSITION

S1 | S 2 | S 3 | L2

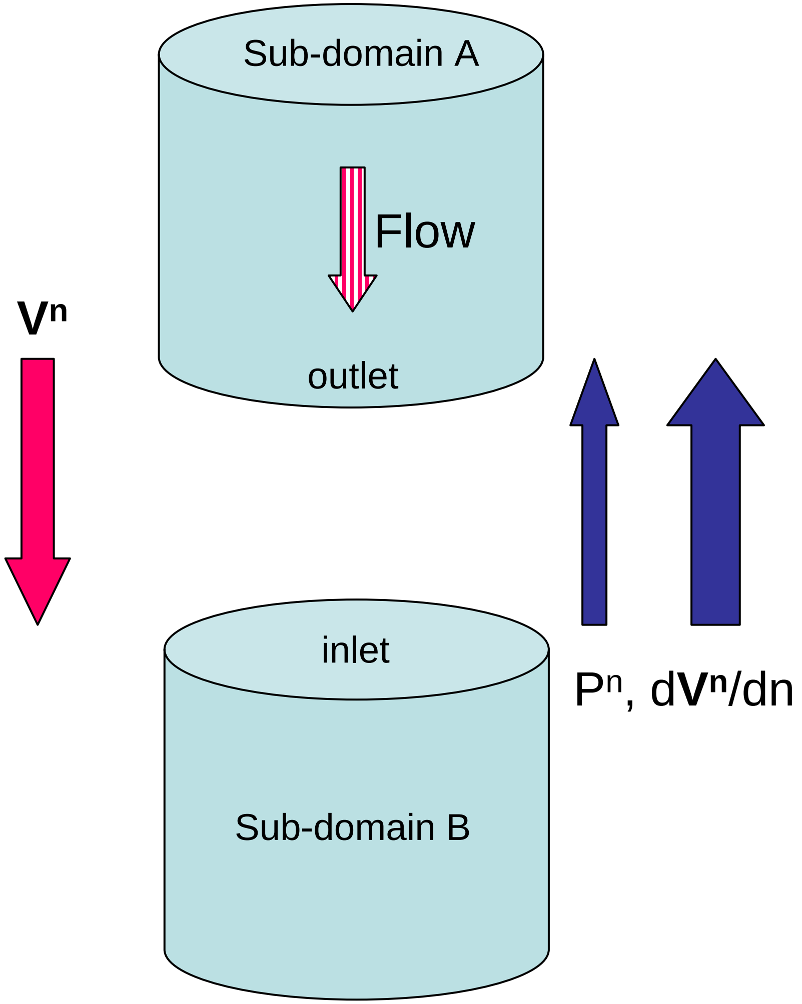
TASK ORIENTED DECOMPOSITION

3D | 3 D | 3 D | 3 D | L3

Two-Level Domain Decomposition: Low Level Communicator Splitting



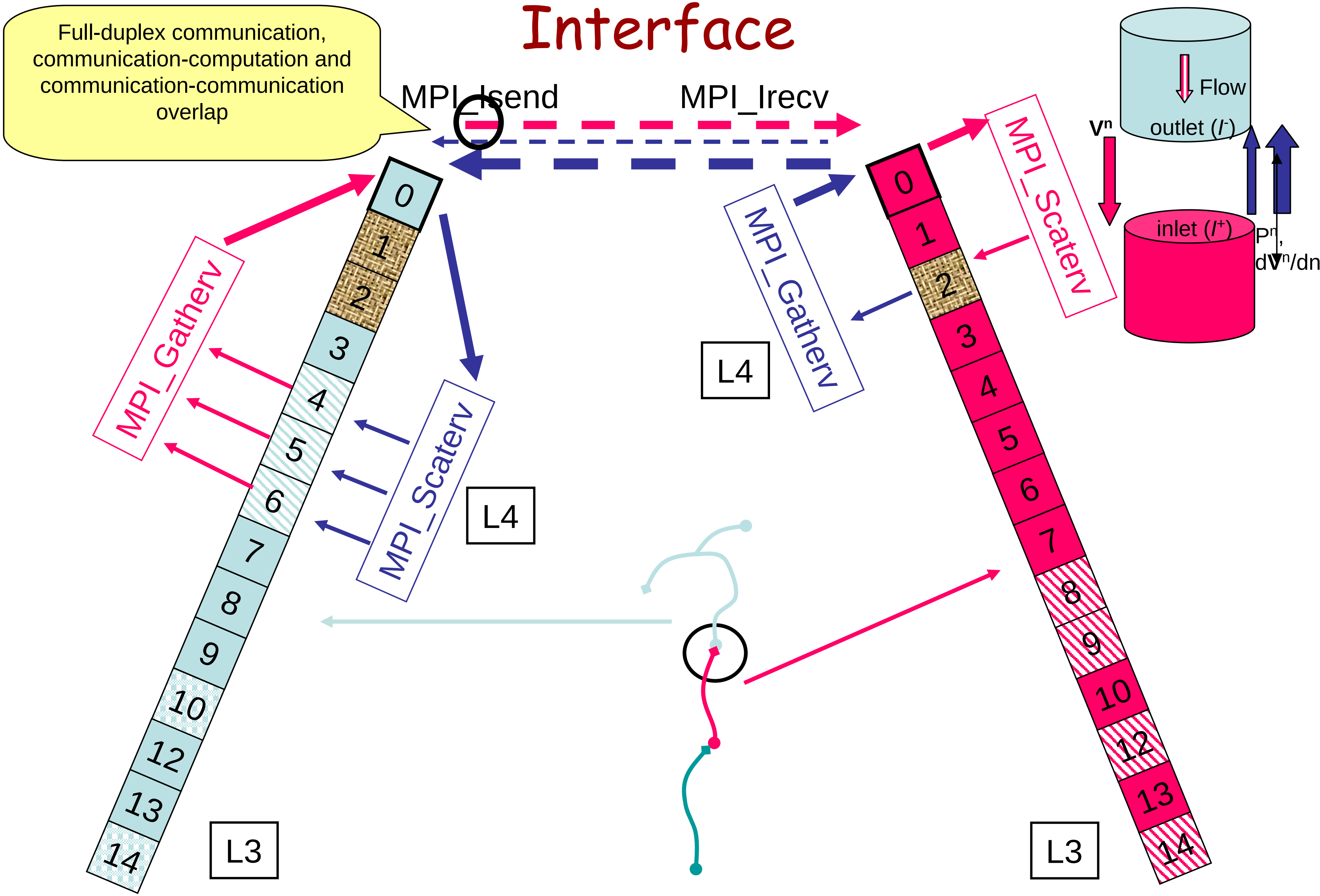
Dual Domain Decomposition: Interface Boundary Conditions



Boundary Condition	Message size
Velocity is computed at outlet and imposed as Dirichlet Boundary Condition at inlet.	N_F*N_M*3* sizeof(double) O(6KB)
Pressure is computed at inlet and imposed as Dirichlet Boundary Condition at outlet.	N_F*N_M* sizeof(double) O(1KB)
Velocity flux from inlet is averaged with velocity flux computed at outlet and imposed as Newman Boundary Condition at outlet.	N_F*(P+3)*$(P+2)^2$ sizeof(double) O(32KB)

N_F, N_M – number of faces and modes,
P – order of polynomial approximation.

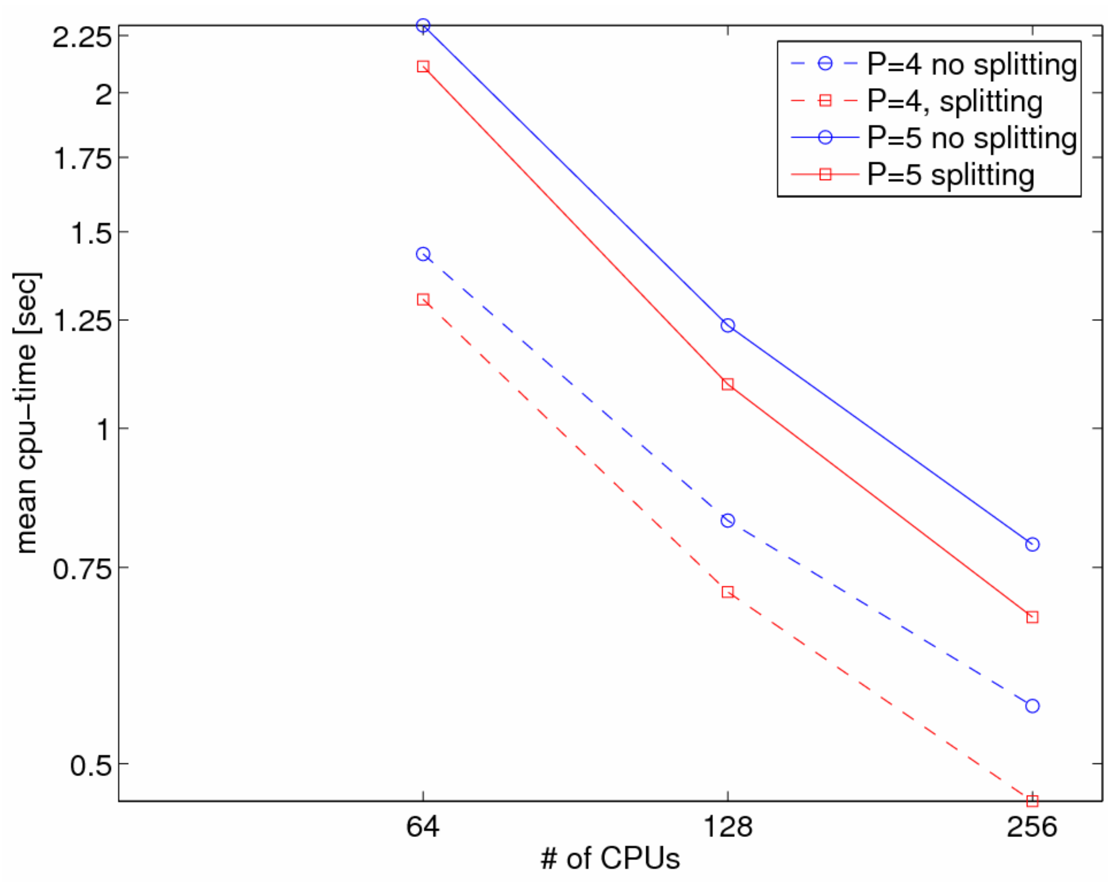
Message Passing Across Sub-Domain Interface



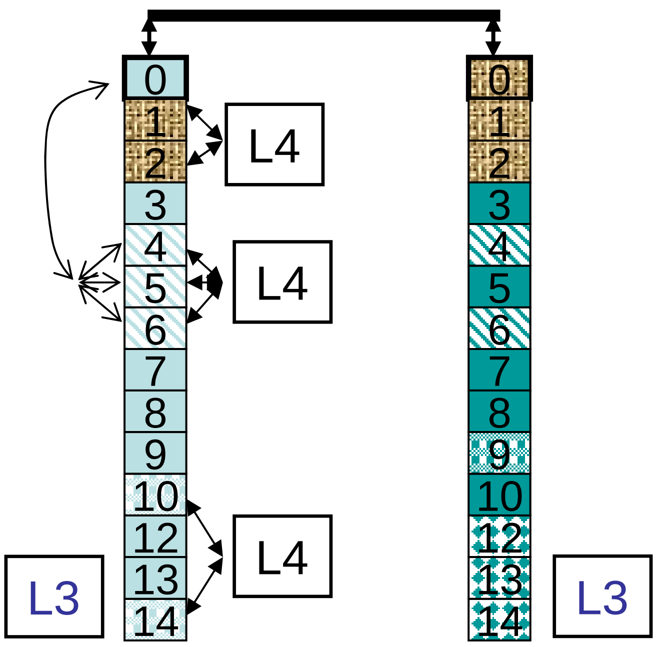
Dual Domain Decomposition Method: details

- Efficiency
- Accuracy

Two Level Domain Decomposition: Efficiency



Mean CPU time required per time step.
Problem size: 67456 tetrahedral elements, polynomial order P=4 and P=5.
Computations were performed on the CRAY XT3 at PSC.



Two Level Domain Decomposition: Accuracy

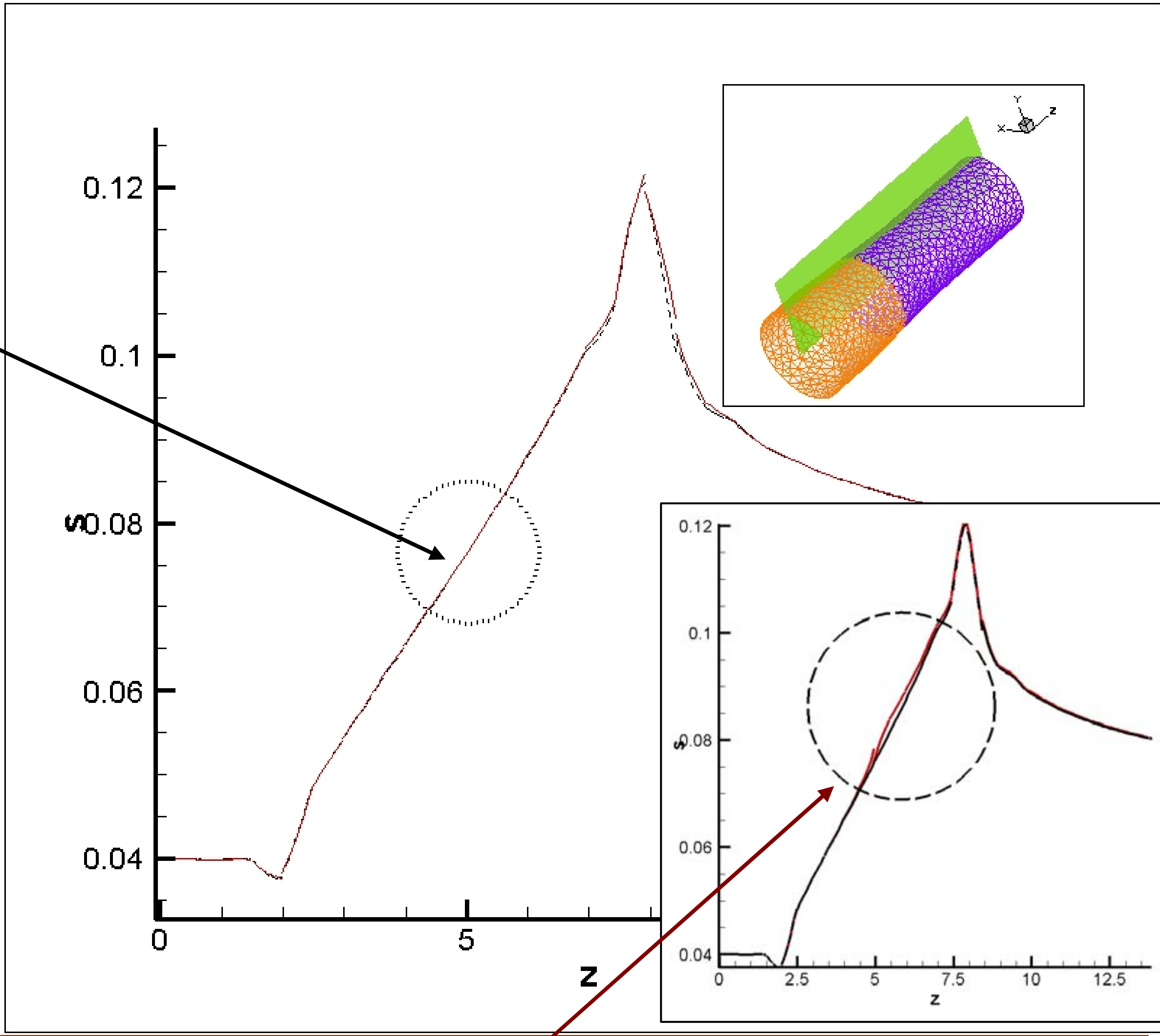
Simulation of steady flow in converging pipe using Dual Domain Decomposition.

Sub-domain interface is at z=5.
Re = 200; P = 5.

B.C. for Velocity are imposed with P=3; for Pressure with P=1 and for Velocity Flux with P=5.

Plot on the right depicts the Wall Shear Stress:

solid line – solution with dual domain decomposition.
dash line – solution with standard domain decomposition.



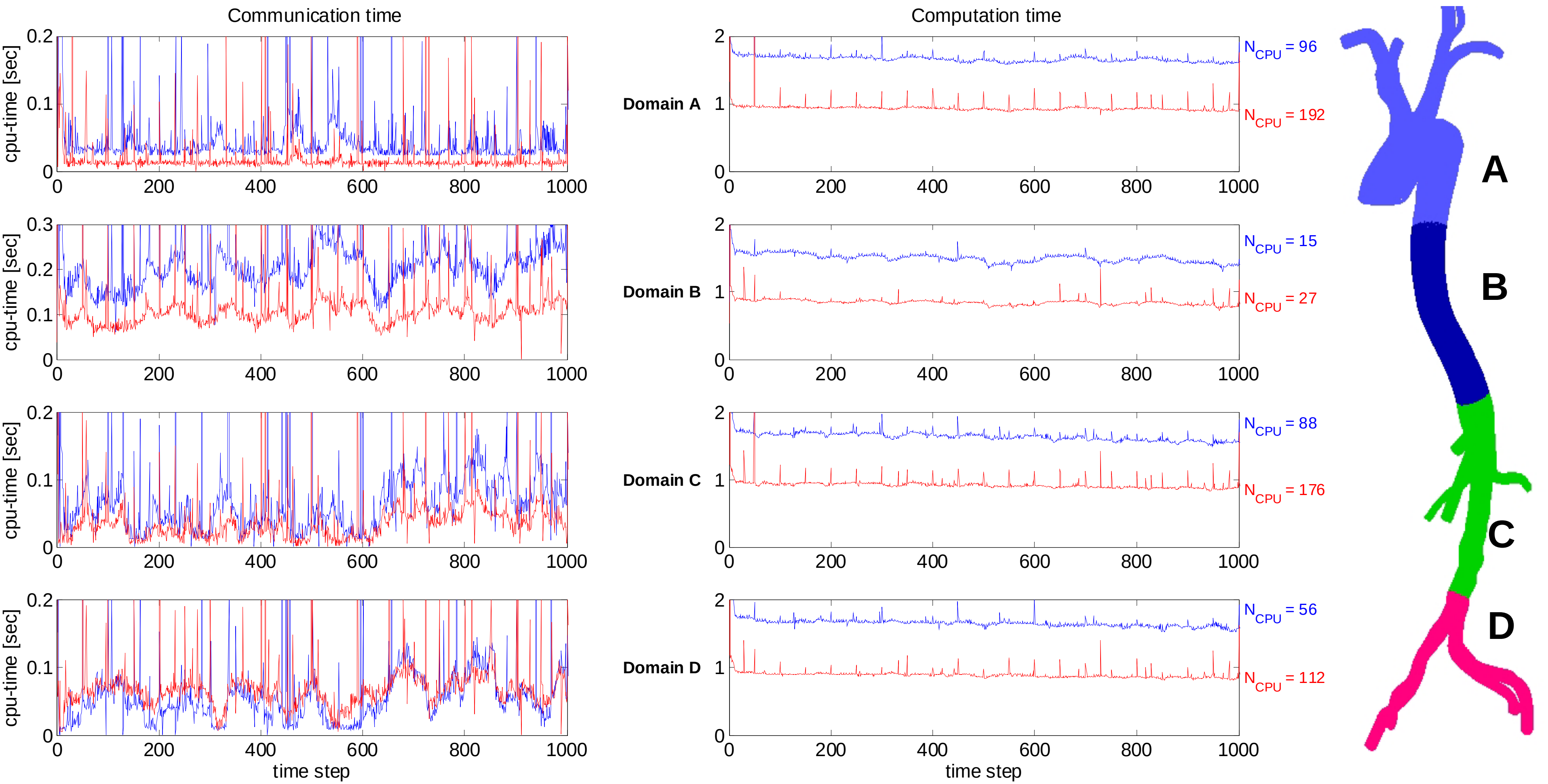
At the inlet of the "blue" sub-domain velocity B.C. are imposed with P=3;

At the outlet of the "orange" sub-domain fully developed boundary conditions are assumed.

Solution of Large Scale Problem with the New Domain Decomposition Method

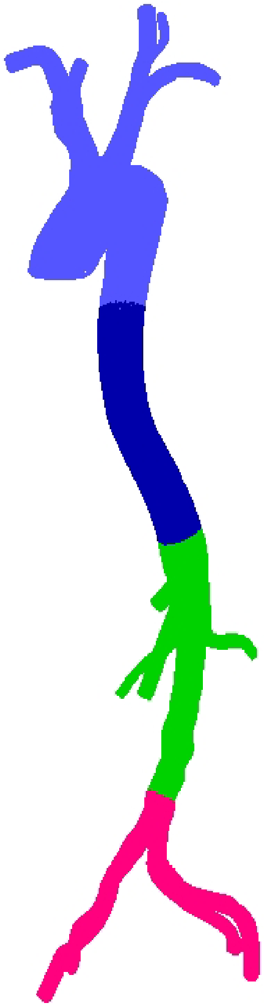
- Numerical Simulation of a Flow in Aorta
- Communication/Computation Time Balance

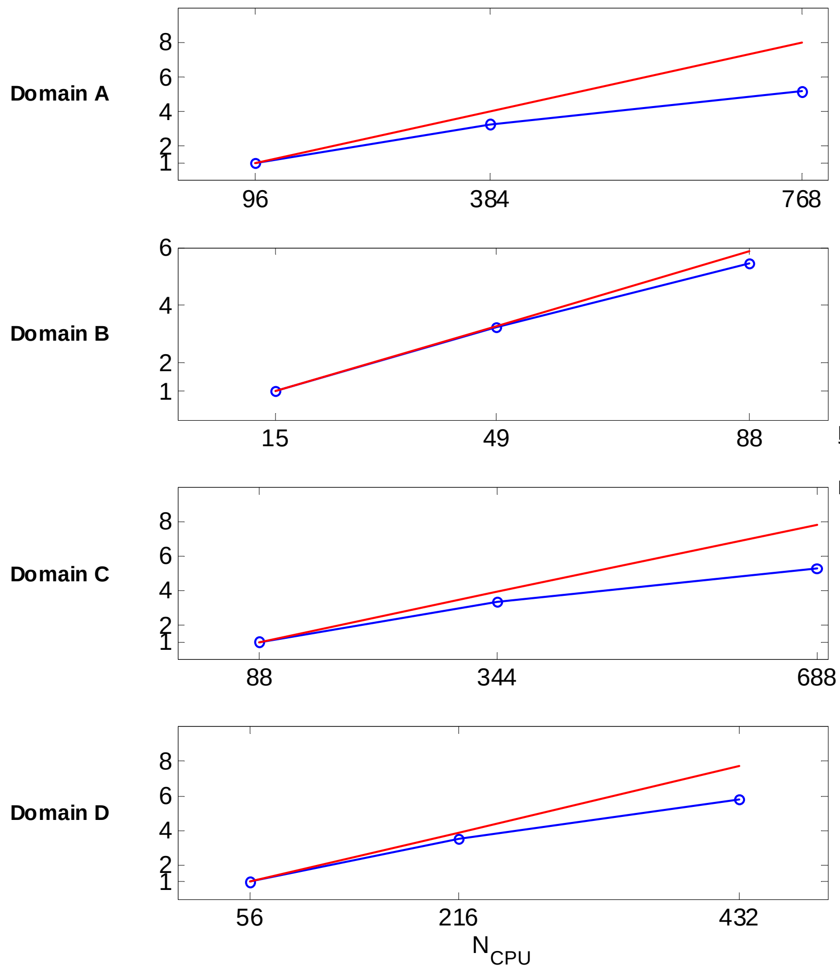
Single machine Computation: Communication / Computation CPU-time balance

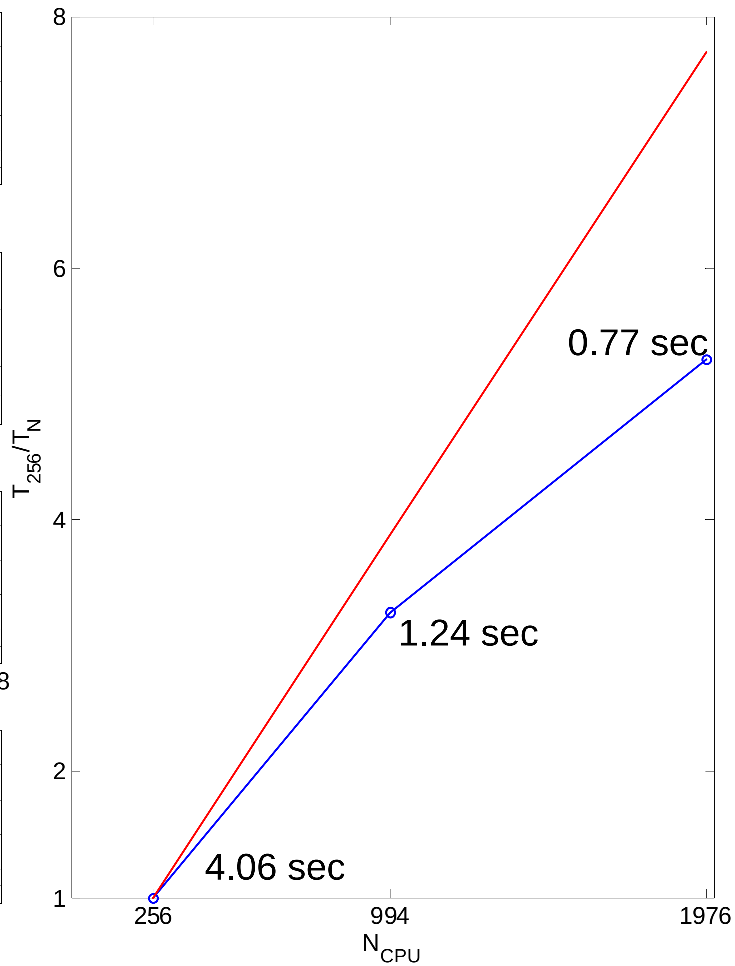


Simulation of a blood flow in Aorta. Nelements = 325,795; P = 6.
Computation was performed on CRAY XT3 with 226 and 508 processors.

Single machine Computation: Standard and Dual Domain Decomposition, Parallel Speed-up

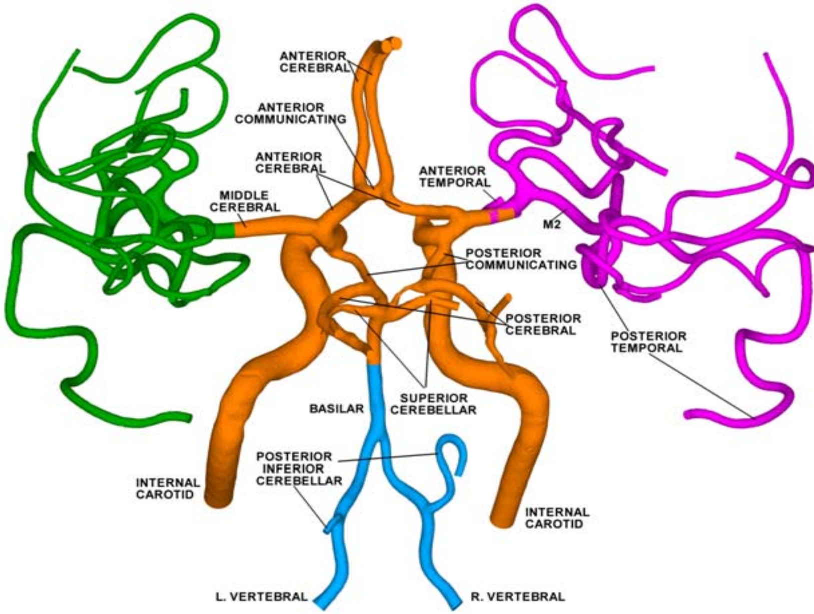


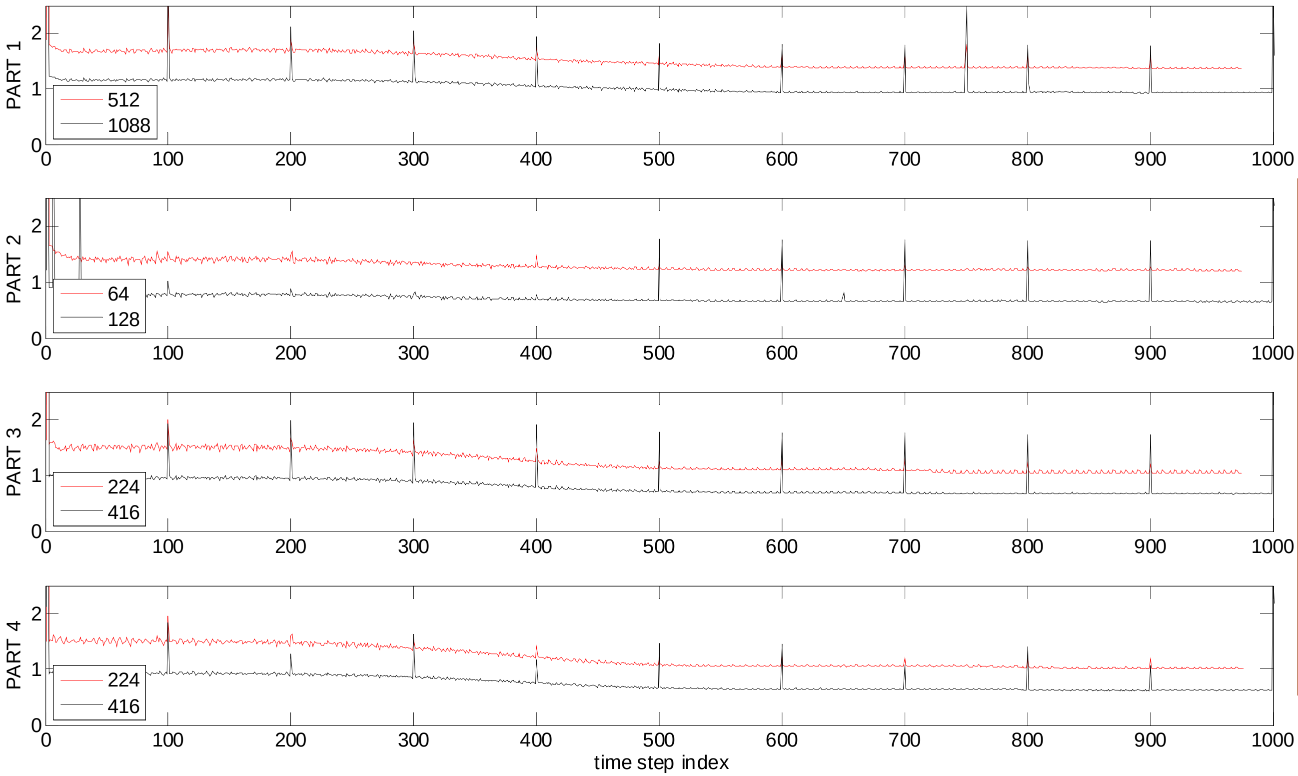




Simulation of a blood flow in Aorta. $N_{elements}$ = 325,795; P = 6. Computation was performed on CRAY XT3.

Single Machine Computation: Coast





Estimated cost of one cardiac cycle simulation on CRAY XT3:

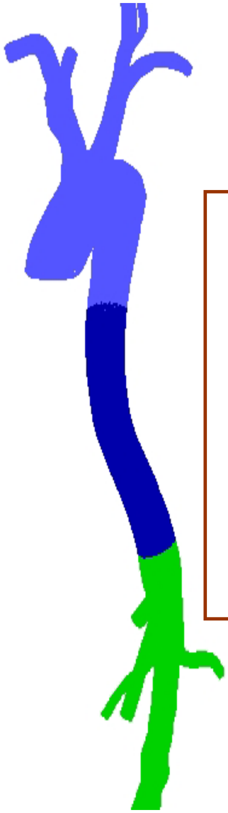
80 hours on 3065 CPUs

120 hours on 2048 CPUs

165 hours on 1024 CPUs

The New Domain Decomposition Method on **TeraGrid**

TeraGrid Cross-Site Computation: Performance

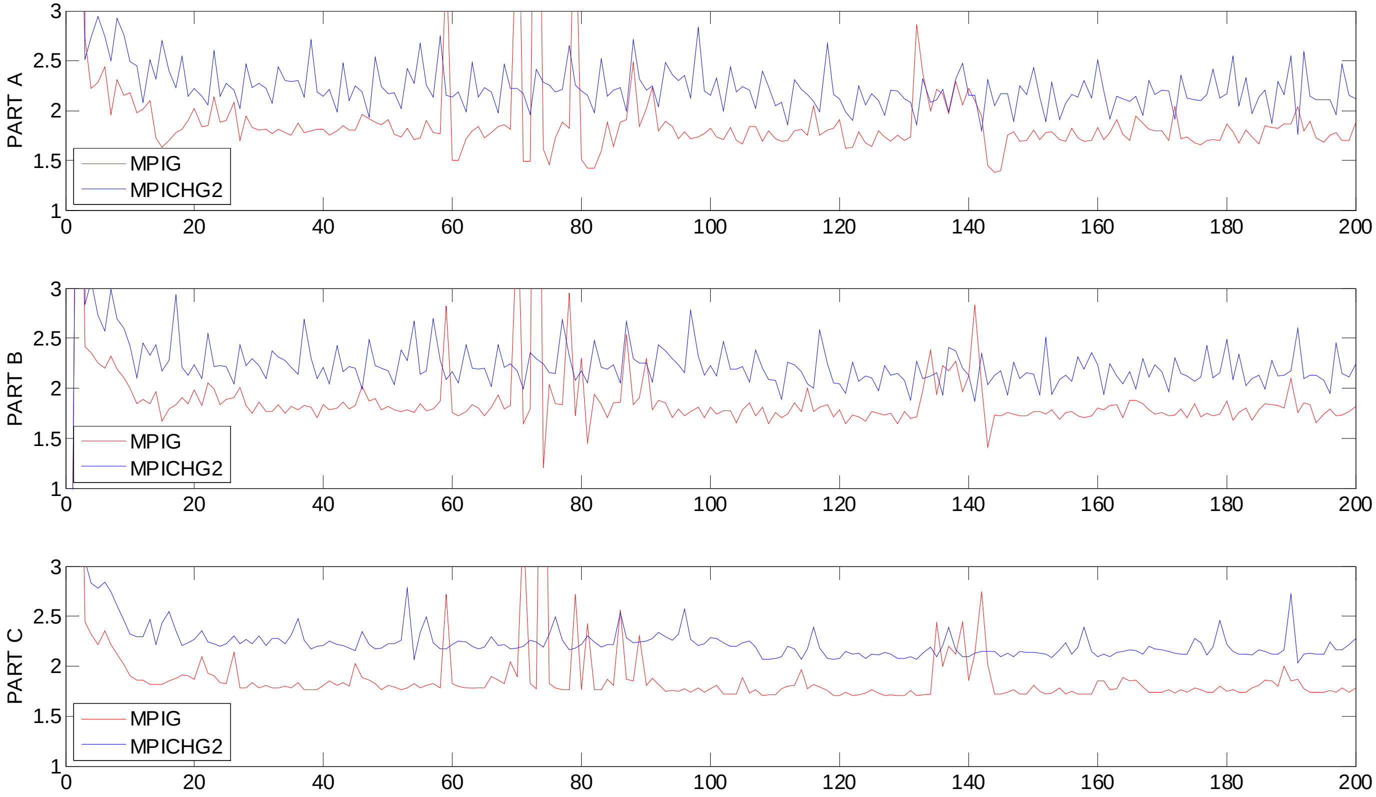


Computational domain:

Composed from 3 PARTS:

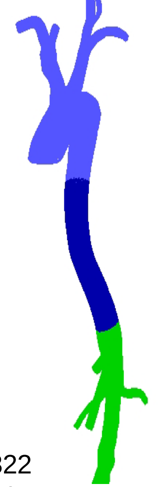
PART A	120,813 elements
PART B	20,797 elements
PART C	106,219 elements

	# of CPUs:	site:
PART A	192	NCSA
PART B	32	NCSA
PART C	128	SDSC



Cpu time measured on rank 0 of each sub-job. Rank 0 handles inter- sub-job communication.

TeraGrid Cross-Site Computation: Performance



NEKTAR-g2 + **MPIg**

- ==> bench_test1.dat <==
- step = 196 COMM_TIME= 0.387799 COMP_TIME= 1.363969
- step = 197 COMM_TIME= 0.431659 COMP_TIME= 1.347054
- step = 198 COMM_TIME= 0.374245 COMP_TIME= 1.333956
- step = 199 COMM_TIME= 0.399128 COMP_TIME= 1.300772
- step = 200 COMM_TIME= 0.465652 COMP_TIME= 1.415958

- ==> bench_test2.dat <==
- step = 196 COMM_TIME= 0.706931 COMP_TIME= 1.088679
- step = 197 COMM_TIME= 0.669795 COMP_TIME= 1.051480
- step = 198 COMM_TIME= 0.692732 COMP_TIME= 1.040437
- step = 199 COMM_TIME= 0.741545 COMP_TIME= 1.024879
- step = 200 COMM_TIME= 0.716007 COMP_TIME= 1.101272

- ==> bench_test3.dat <==
- step = 196 COMM_TIME= **0.021616** COMP_TIME= 1.737263
- step = 197 COMM_TIME= **0.021568** COMP_TIME= 1.722539
- step = 198 COMM_TIME= **0.021713** COMP_TIME= 1.760799
- step = 199 COMM_TIME= **0.021614** COMP_TIME= 1.718854
- step = 200 COMM_TIME= **0.021682** COMP_TIME= 1.760359

NEKTAR-g2 + **MPICHG2**

- ==> bench_test1.dat <==
- step = 196 COMM_TIME= 0.533426 COMP_TIME= 1.575322
- step = 197 COMM_TIME= 0.374780 COMP_TIME= 1.590261
- step = 198 COMM_TIME= 0.857370 COMP_TIME= 1.605632
- step = 199 COMM_TIME= 0.548640 COMP_TIME= 1.605805
- step = 200 COMM_TIME= 0.513864 COMP_TIME= 1.599325

- ==> bench_test2.dat <==
- step = 196 COMM_TIME= 0.889990 COMP_TIME= 1.060204
- step = 197 COMM_TIME= 1.369228 COMP_TIME= 1.078424
- step = 198 COMM_TIME= 1.081203 COMP_TIME= 1.073181
- step = 199 COMM_TIME= 1.083127 COMP_TIME= 1.036677
- step = 200 COMM_TIME= 1.182271 COMP_TIME= 1.059155

- ==> bench_test3.dat <==
- step = 196 COMM_TIME= **0.121957** COMP_TIME= 2.123489
- step = 197 COMM_TIME= **0.122057** COMP_TIME= 2.040032
- step = 198 COMM_TIME= **0.122080** COMP_TIME= 2.034881
- step = 199 COMM_TIME= **0.122216** COMP_TIME= 2.091826
- step = 200 COMM_TIME= **0.122080** COMP_TIME= 2.158722

Cpu time is measured (in seconds) on rank 0 of each sub-job. Communication time includes extra time we need to create message (MPI_Gatherv) pass it to partner cpu from another subjob and then scatter (with MPI_Scatterv) within appropriate group of processors.

Summary

- We developed and implemented a new scalable approach for solution of large problems on the TeraGrid and beyond.
- Overlapping computation with cross-site communication, performing simultaneous communication and full-duplex communication over multiple channels hides the expensive inter-site latency on TeraGrid.
- Future plan: improve communication algorithm (between different process groups).

Future plans

We aim to establish a biomechanics gateway on the TeraGrid and make the arterial tree a platform and a simulation framework for further developments and systematic studies in hemodynamics, disease modeling, and drug delivery.

http://www.cfm.brown.edu/crunch/ATREE/index.html

CRUNCH GROUP

Thank you!