New CPU, GPU, and Big Data Clusters in the UMBC High Performance Computing Facility

Carlos Barajas, Matthias K. Gobbert

UMBC High Performance Computing Facility (HPCF)
hpcf.umbc.edu
and
Department of Mathematics and Statistics
University of Maryland, Baltimore County (UMBC)

Acknowledgments: NSF, UMBC, HPCF, CIRC, Math & Stat

The UMBC High Performance Computing Facility (HPCF)

UMBC High Performance Computing Facility (HPCF)

The community-based, interdisciplinary core facility for scientific computing and research on parallel algorithms at UMBC

- Initiated in 2008 with participation of over 20 faculty from more than 10 departments and research centers from all three colleges at UMBC
- Supported by UMBC seed funding in 2008, contributions from faculty, and MRI grants from the National Science Foundation ($200k in 2008, $300k in 2012, $550k in 2017)
- Major purchases in 2008 ($270k), 2009 ($600k), 2013 ($540k), and 2018 ($790k)
- Grants based on concept of need for (i) hardware (→ NSF grants), (ii) system administration (→ institutional support), (iii) user support (→ consulting approach), and (iv) appropriate usage policies (→ governance committee).
- System administration by Division of Information Technology; user support by HPCF RAs in collaboration with CIRC
- Governed by HPCF Governance Committee; point of contact: Matthias K. Gobbert, gobbert@umbc.edu, 410–455–2404
- Over 400 users since 2008, in one month 100 users from 30 groups.
- Over 250 publications, incl. over 100 journal papers
- All details, list of projects and publications, user info at `hpcf.umbc.edu`.

HPCF Governance Committees (as of 11/20/18)

- HPCF Governance Committee:
  - Ruben Delgado (Joint Center for Earth Systems Technology)
  - Don Engel (Office of the Vice President for Research)
  - Matthias K. Gobbert (Mathematics and Statistics, Chair)
  - Curtis R. Menyuk (Computer Science and Electrical Engineering)
  - Marc Olano (Computer Science and Electrical Engineering)
  - Larrabee Strow (Physics)
  - Jianwu Wang (Information Systems)
  - Claire Welty (Chemical, Biochemical & Environmental Engineering and CUERE)
  - Meilin Yu (Mechanical Engineering)
  - Zhibo Zhang (Physics)
- Former members of the governance committee:
  - Lynn Sparling (Physics)
  - Ian Thorpe (Chemistry and Biochemistry)
- They represent over 400 long-term research users from over 35 research groups.

taki 2018 Addition

With Matthias Gobbert and HPCF RA Carlos Barajas

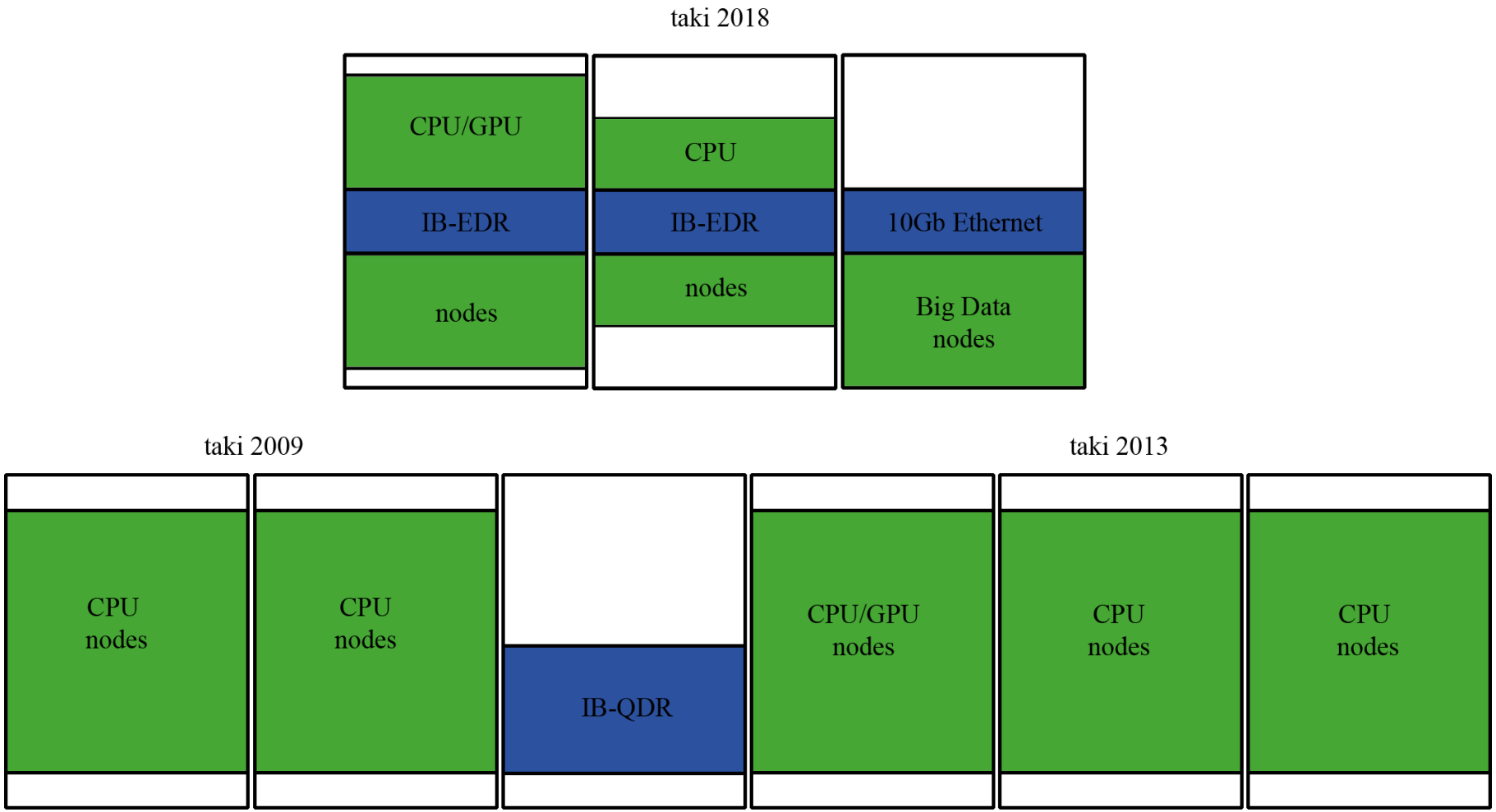
Structure of the taki Clusters in HPCF by Purchase

- HPCF2018:
  - 42 compute and 2 develop nodes (each two 18-core Skylake CPUs)
  - 1 GPU node (4 NVIDIA Tesla V100 GPUs connected by NVLink)
  - 8 Big Data nodes (each two Skylake CPUs and 48 TB disk space)
  - 2 login/user nodes, 1 management node
- HPCF2013:
  - 49 compute and 2 develop nodes (each two 8-core Ivy Bridge CPUs)
  - 18 CPU/GPU nodes (hybrid nodes with 2 CPUs and 2 GPUs)
  - 2 login/user nodes, 1 management node
- HPCF2009:
  - 82 compute and 2 develop nodes (each two 4-core Nehalem CPUs)
  - 1 login/user node, 1 management node

Being removed from service now:

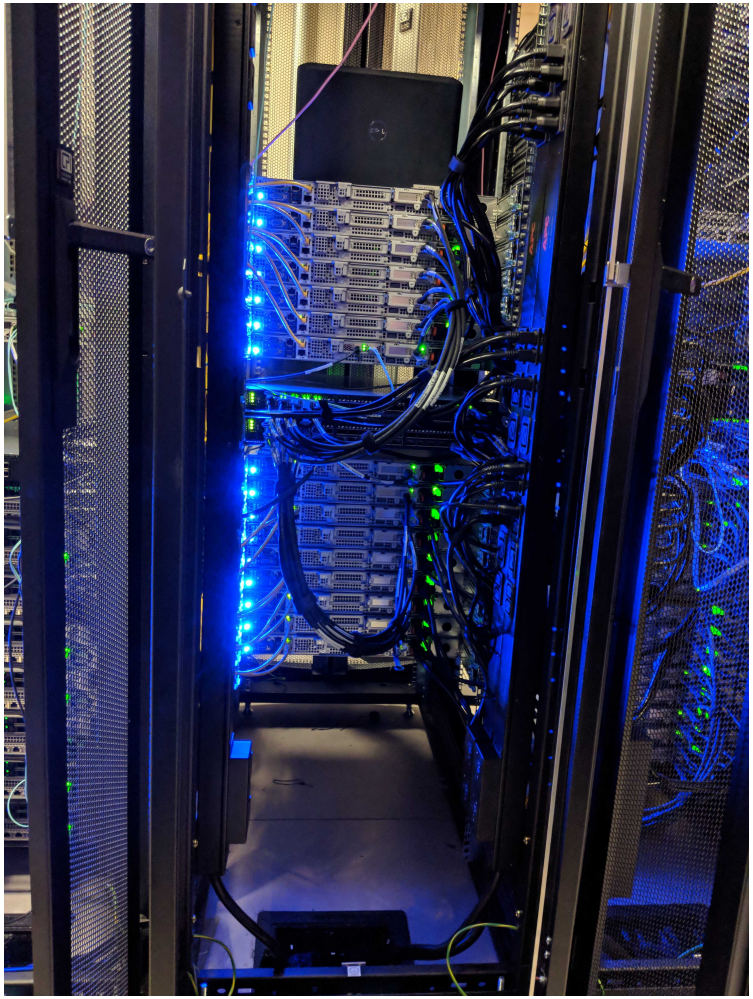
- HPCF2010 (gift from NASA):
  - 162 compute and 2 develop nodes (each two 4-core Nehalem CPUs)

Taki: System Layout

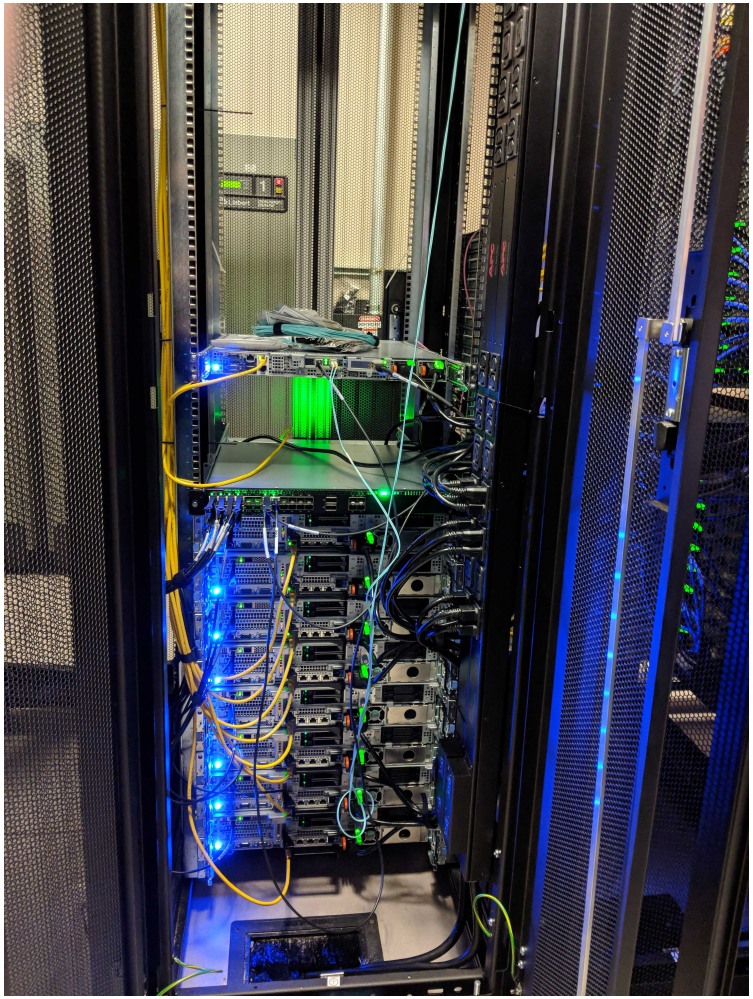


HPCF2018: First CPU Rack Front and Back

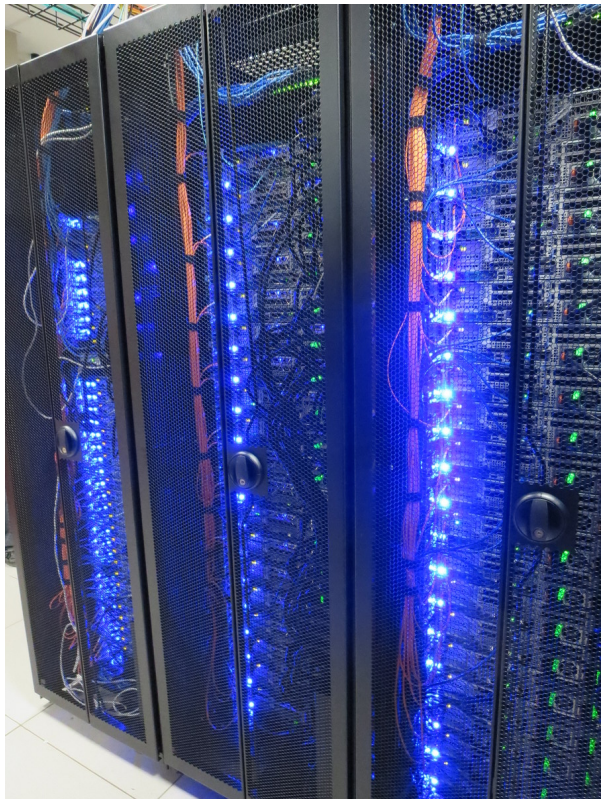
HPCF2018: Second CPU Rack Front and Back



HPCF2018: Big Data Rack Front and Back



HPCF2013: Racks Front and Back



HPCF2009: iDataPlex Rack and Detail

HPCF2009 and HPCF2013: Front and Back of the QDR IB Switch

CPU, GPU, and Big Data Clusters in taki

- CPU Cluster: 179 nodes total to over 3000 cores and over 20 TB memory!
  - HPCF2018: 42 compute and 2 develop nodes
    (each two 18-core Skylake CPUs and 384 GB memory)
  - HPCF2013: 49 compute and 2 develop nodes
    (each two 8-core Ivy Bridge CPUs and 64 GB memory)
  - HPCF2009: 82 compute and 2 develop nodes
    (each two 4-core Nehalem CPUs and 24 GB memory)
- GPU Cluster:
  - HPCF2018: 1 GPU node
    (4 NVIDIA Tesla V100 GPUs connected by NVLink, 2 CPUs)
  - HPCF2013: 18 CPU/GPU nodes
    (hybrid nodes with 2 CPUs and 2 NVIDIA K20 GPUs)
- Big Data Cluster:
  - HPCF2018: 8 Big Data nodes (each 2 CPUs and 48 TB disk space)
- Other Nodes:
  - 2 login/user nodes (taki-usr1, taki-usr2)
  - 1 management node

Details of the CPU Cluster in taki

HPCF2018, 42 compute nodes total to 1512 cores and over 15 TB pooled memory:

- 42 compute (cnode002–029,031–044) and 2 develop nodes (cnode001,030), each with two 18-core Intel Xeon Gold 6140 Skylake CPUs (2.3 GHz clock speed, 24.75 MB L3 cache, 6 memory channels, 140 W power), 384 GB memory (12×32 GB DDR4), and a 120 GB SSD drive;
- connected by a network of four 36-port EDR (Enhanced Data Rate) InfiniBand switches (100 Gb/s bandwidth, 90 ns latency);

HPCF2013, 49 compute nodes total to 784 cores and over 3 TB of pooled memory:

- 49 compute and 2 develop nodes, each with two 8-core Intel E5-2650v2 Ivy Bridge CPUs (2.6 GHz clock speed, 20 MB L3 cache, 4 memory channels), 64 GB memory (8×8 GB DDR3), and a 500 GB local hard drive,
- connected by a QDR (quad-data rate) InfiniBand switch;

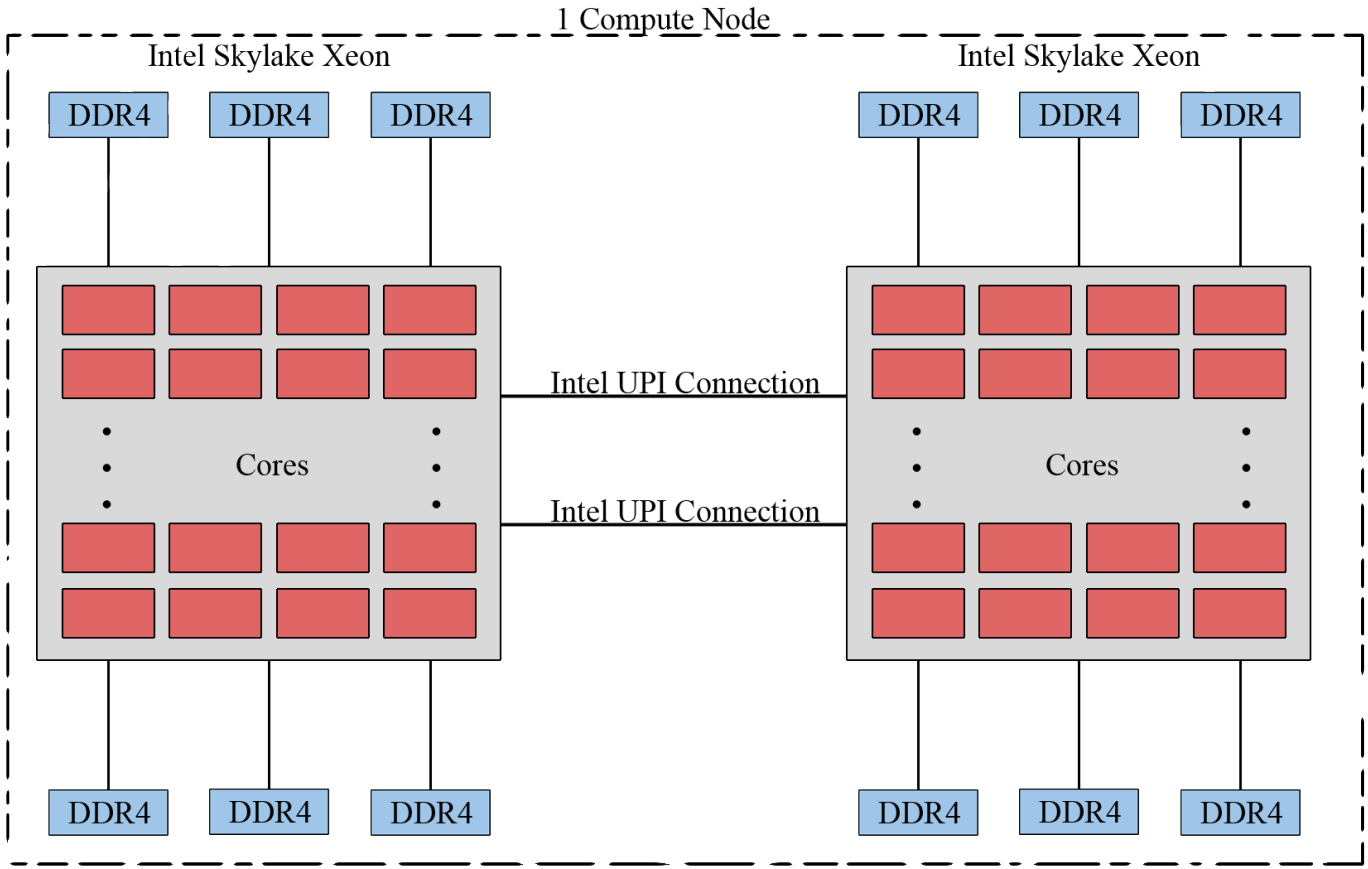
HPCF2009, 82 compute nodes total to 656 cores and 2 TB of pooled memory:

- 82 compute and 2 develop nodes, each with two 4-core Intel Nehalem X5550 CPUs (2.6 GHz clock speed, 8 MB L3 cache, 3 memory channels), 24 GB memory (6×4 GB DDR3), and a 120 GB local hard drive,
- connected by a QDR (quad-data rate) InfiniBand switch.

HPCF2018: Schematic of Three Racks with Node Names

RU	A6	RU	A5	RU	A4
42	Management GbE Switch	42	Management GbE Switch	42	
41		41		41	
40		40		40	
39		39		39	
38	cnode029	38		38	
37	cnode028	37		37	
36	cnode027	36		36	
35	cnode026	35		35	
34	cnode025	34		34	
33	cnode024	33		33	
32	cnode023	32		32	
31	cnode022	31		31	
30	cnode021	30		30	
29	cnode020	29		29	
28	cnode019	28	cnode044	28	
27	gpunode001	27	cnode043	27	
26	cnode018	26	cnode042	26	
25	cnode017	25	cnode041	25	
24	cnode016	24	cnode040	24	
23	cnode015	23	cnode039	23	
22	hnode001	22	login001	22	login002
21		21		21	
20	Mellanox IB Switch Leaf 3	20	Mellanox IB Switch Leaf 1	20	
19	Mellanox IB Switch Core	19	Mellanox IB Switch Leaf 2	19	
18		18		18	
17	cnode014	17	cnode038	17	BigData 10GbE Switch
16	cnode013	16	cnode037	16	bdnod008
15	cnode012	15	cnode036	15	
14	cnode011	14	cnode035	14	bdnode007
13	cnode010	13	cnode034	13	
12	cnode009	12	cnode033	12	bdnod006
11	cnode008	11	cnode032	11	
10	cnode007	10	cnode031	10	bdnode005
9	cnode006	9	cnode030	9	
8	cnode005	8		8	bdnod004
7	cnode004	7		7	
6	cnode003	6		6	bdnode003
5	cnode002	5		5	
4	cnode001	4		4	bdnod002
3		3		3	
2		2		2	bdnode001
1		1		1	

Dual-Socket Compute Node with Two 18-Core Intel Skylake CPUs



- two 18-core Intel Xeon Gold 6140 Skylake CPUs
  (2.3 GHz clock speed, 24.75 MB cache, 6 memory channels, 140 W power),
- 384 GB memory (12×32 GB DDR4)

Performance Studies for the Poisson Equation on taki

Classical Elliptic Test Problem for Numerical PDE Methods

Poisson equation with homogeneous Dirichlet boundary conditions in 2-D:
Find the solution $u(x, y)$ on the unit square $\Omega = (0, 1) \times (0, 1) \subset \mathbb{R}^2$ such that

$$-\triangle u = f \quad \text{in } \Omega, \qquad \triangle u = u_{xx} + u_{yy} \text{ Laplace operator}$$
$$u = 0 \quad \text{on } \partial\Omega \qquad f(x, y) \text{ given function}$$

On mesh $\Omega_h = \{(x_i, y_j) = (ih, jh), i, j = 0, \ldots, N + 1\}$ with $h = \frac{1}{N+1}$,
approximation by the **finite difference method**

$$-\triangle u(x_i, y_j) \approx \frac{-u_{i-1,j} + 2u_{i,j} - u_{i+1,j}}{h^2} + \frac{-u_{i,j-1} + 2u_{i,j} - u_{i,j+1}}{h^2}$$

defines approximations $u_{ij} \approx u(x_i, y_j)$ for $i, j = 1, \ldots, N$ on the interior of $\Omega_h \subset \overline{\Omega}$.
Numerical analysis predicts convergence of the numerical solution $u_h(x, y)$ as

$$\|u - u_h\|_{L^\infty(\Omega)} \leq C\, h^2 \quad \text{as } h \to 0$$

with constant C independent of h in the function norm
$\|u - u_h\| := \|u - u_h\|_{L^\infty(\Omega)} := \sup_{(x,y)\in\Omega_h} |u(x, y) - u_h(x, y)|$.
For sufficiently small h, convergence is indicated if the ratio of errors on consecutively refined meshes behaves like

$$\text{Ratio} = \frac{\|u - u_{2h}\|}{\|u - u_h\|} \approx \frac{C\,(2h)^2}{C\,h^2} = 4 \tag{1}$$

Classical Elliptic Test Problem for Numerical Linear Algebra

The finite difference discretization results in a system of linear equations $A\,\mathbf{u} = \mathbf{b}$ for the unknowns $\mathbf{u} = (\mathbf{u}_k) = (u_{ij})$ with ordering $k = i + N(j-1)$ for $i, j = 1, \ldots, N$, and analogously $\mathbf{b} = (\mathbf{b}_k) = (h^2\, f(x_i, y_j)$. The system matrix

$$A = \begin{bmatrix} S & T & & & \\ T & S & T & & \\ & \ddots & \ddots & \ddots & \\ & & T & S & T \\ & & & T & S \end{bmatrix} \in \mathbb{R}^{N^2 \times N^2} \qquad \begin{aligned} S &= \text{tridiag}(-1, 4, -1) \in \mathbb{R}^{N \times N}, \\ T &= -I \in \mathbb{R}^{N \times N} \end{aligned}$$

is large, sparse, highly structured, and symmetric positive definite.
Therefore, optimal iterative method in the family of Krylov subspace methods is the **conjugate gradient (CG) method**. Methods in this family only require the matrix-vector product $\mathbf{v} = A\,\mathbf{u}$ (and not A itself), hence a matrix-free implementation is possible that saves memory. $\Longrightarrow$ CG requires

- storage: 4 vectors (solution $\mathbf{u}$, residual $\mathbf{r}$, search direction $\mathbf{p}$, auxiliary $\mathbf{q}$)
- operations: 1 matrix-vector prod., 2 inner prod., 3 vector updates per iteration
- communications among parallel processes:
  2 `MPI_Allreduce`, 2 `MPI_Isend/MPI_Irecv` per iteration

Elliptic Test Problem

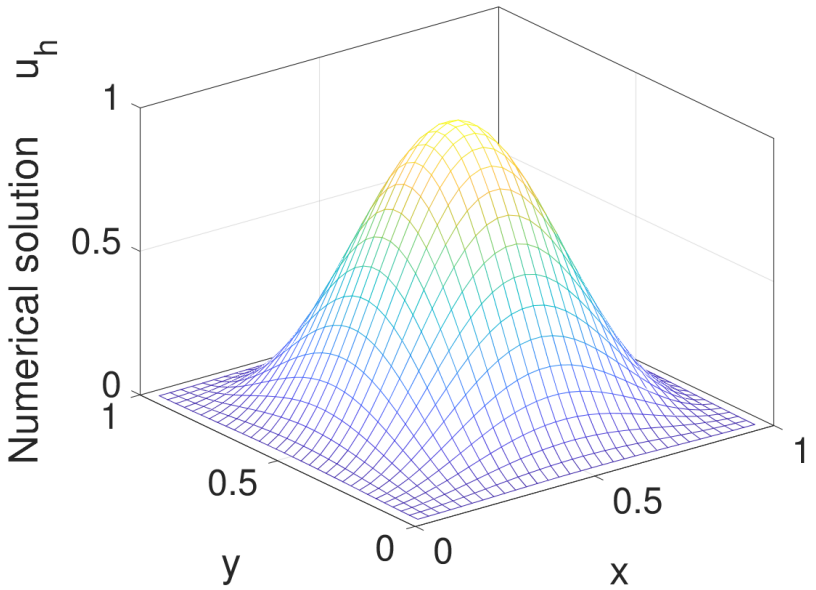
Test problem has right-hand side function

$$f(x,y) = (-2\pi^2)\left(\cos(2\pi x)\,\sin^2(\pi y) + \sin^2(\pi x)\,\cos(2\pi y)\right),$$

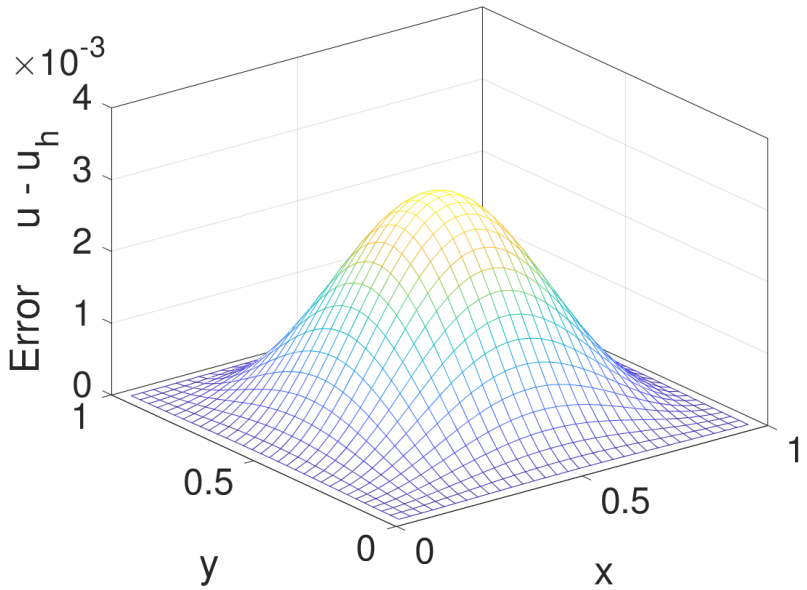
This has a known solution

$$u(x,y) = \sin^2(\pi x)\,\sin^2(\pi y)$$

On a mesh with 32×32 points and mesh spacing $h = 1/33 = 0.030303$, the numerical solution $u_h(x,y)$ and its error $u - u_h$ can be plotted as



(a) Numerical solution u_h



(b) Error $u - u_h$

taki 2018: convergence study with serial code except as noted

Poisson equation in 2-D, finite difference discretization, conjugate gradient method, matrix-free implementation, mesh resolution $N \times N$, system dimension DOF $= N^2$, FD error and ratio, CG iterations, observed memory usage in GB [HPCF–2018–18]

N	DOF	$\|u - u_h\|$	Ratio	iter	Time	Memory
32	1,024	3.0128e–03	—	48	00:00:00	<1 GB
64	4,096	7.7811e–04	3.87	96	00:00:00	<1 GB
128	16,384	1.9765e–04	3.94	192	00:00:00	<1 GB
256	65,536	4.9797e–05	3.97	387	00:00:00	<1 GB
512	262,144	1.2494e–05	3.99	783	00:00:01	<1 GB
1024	1,048,576	3.1266e–06	4.00	1,581	00:00:09	<1 GB
2048	4,194,304	7.8019e–07	4.01	3,192	00:01:34	<1 GB
4096	16,777,216	1.9366e–07	4.03	6,452	00:13:32	<1 GB
8192	67,108,864	4.7392e–08	4.09	13,033	01:52:32	2.02 GB
16384	268,435,456	1.1647e–08	4.07	26,316	15:07:31	8.02 GB
32768	1,073,741,824	2.6004e–09	4.48	53,141	129:18:17	32.02 GB
*65536	4,294,967,296	8.9963e–10	2.89	107,261	*02:44:12	*161.53 GB
*131072	17,179,869,184	3.0435e–10	2.96	216,433	*23:01:28	*547.54 GB

*This case uses 32 cores on 32 nodes; memory is the total over all 1024 processes.
$N = 32768$ using 32 cores on 32 nodes: 00:23:38.

taki 2018: wall clock time in HH:MM:SS of MPI-only code

Mesh resolution $N \times N = 4096 \times 4096$, system dimension 16777216						
	1 node	2 nodes	4 nodes	8 nodes	16 nodes	32 nodes
1 proc per node	00:13:32	00:06:36	00:03:12	00:01:36	00:00:38	00:00:14
2 proc per node	00:06:41	00:03:12	00:01:30	00:00:37	00:00:14	00:00:07
4 proc per node	00:03:23	00:01:38	00:00:46	00:00:20	00:00:07	00:00:04
8 proc per node	00:01:48	00:00:51	00:00:24	00:00:10	00:00:04	00:00:02
16 proc per node	00:01:15	00:00:38	00:00:19	00:00:06	00:00:02	00:00:01
32 proc per node	00:01:03	00:00:31	00:00:14	00:00:04	00:00:02	00:00:02
Mesh resolution $N \times N = 8192 \times 8192$, system dimension 67108864						
	1 node	2 nodes	4 nodes	8 nodes	16 nodes	32 nodes
1 proc per node	01:52:32	00:56:47	00:28:14	00:14:47	00:07:20	00:03:40
2 proc per node	00:57:03	00:28:08	00:13:49	00:06:38	00:03:04	00:01:21
4 proc per node	00:29:10	00:14:27	00:07:10	00:03:27	00:01:36	00:00:42
8 proc per node	00:15:12	00:07:34	00:03:44	00:01:49	00:00:51	00:00:22
16 proc per node	00:10:09	00:05:12	00:02:39	00:01:17	00:00:39	00:00:15
32 proc per node	00:08:45	00:04:24	00:02:11	00:01:05	00:00:31	00:00:11

In 2003, 4096×4096 was the largest mesh possible in 1 GB of memory per node.

taki 2018: wall clock time in HH:MM:SS of MPI-only code

Mesh resolution $N \times N = 16384 \times 16384$, system dimension 268435456						
	1 node	2 nodes	4 nodes	8 nodes	16 nodes	32 nodes
1 proc per node	15:07:31	07:34:50	03:50:43	01:55:51	00:58:18	00:29:45
2 proc per node	07:37:16	03:50:04	01:54:52	00:56:46	00:27:51	00:13:26
4 proc per node	03:56:04	01:57:46	00:58:49	00:29:18	00:14:26	00:06:57
8 proc per node	02:03:10	01:01:35	00:30:48	00:15:22	00:07:35	00:03:44
16 proc per node	01:22:35	00:41:06	00:20:37	00:10:35	00:05:16	00:02:37
32 proc per node	01:11:27	00:35:42	00:17:55	00:09:02	00:04:40	00:02:21

Mesh resolution $N \times N = 32768 \times 32768$, system dimension 1073741824						
	1 node	2 nodes	4 nodes	8 nodes	16 nodes	32 nodes
1 proc per node	129:18:17	64:15:32	32:08:45	16:13:22	08:06:37	04:01:54
2 proc per node	64:44:32	32:21:54	16:12:47	09:35:55	04:21:39	02:39:27
4 proc per node	33:04:35	18:11:46	09:54:45	04:58:51	02:16:26	01:20:41
8 proc per node	17:13:52	08:44:30	04:36:02	02:14:35	01:17:27	00:40:58
16 proc per node	11:19:13	05:53:54	03:00:33	01:29:34	00:45:11	00:24:23
32 proc per node	09:40:51	04:52:29	02:27:33	01:14:26	00:38:57	00:20:38

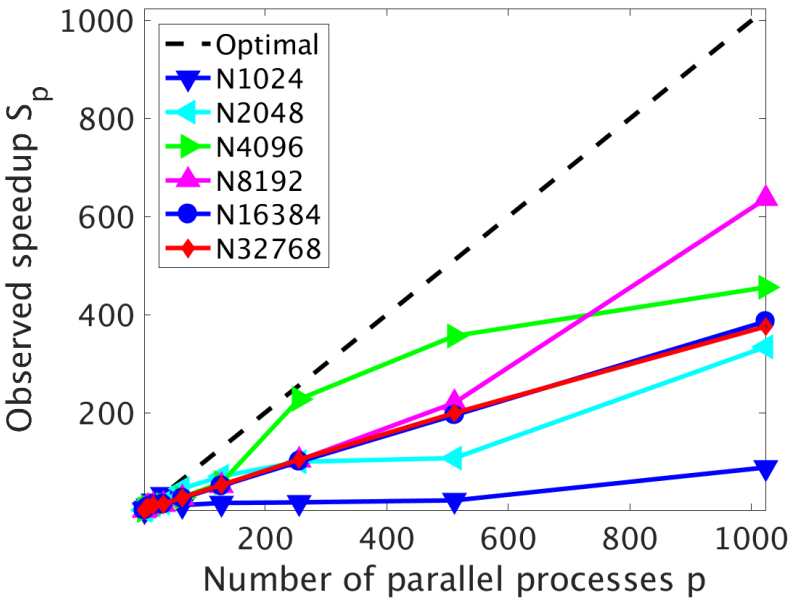
- 65536×65536 mesh using 32 cores on 32 nodes: 02:44:12
- 131072×131072 mesh using 32 cores on 32 nodes: 23:01:28

taki 2018: strong scalability by p processes of MPI-only code

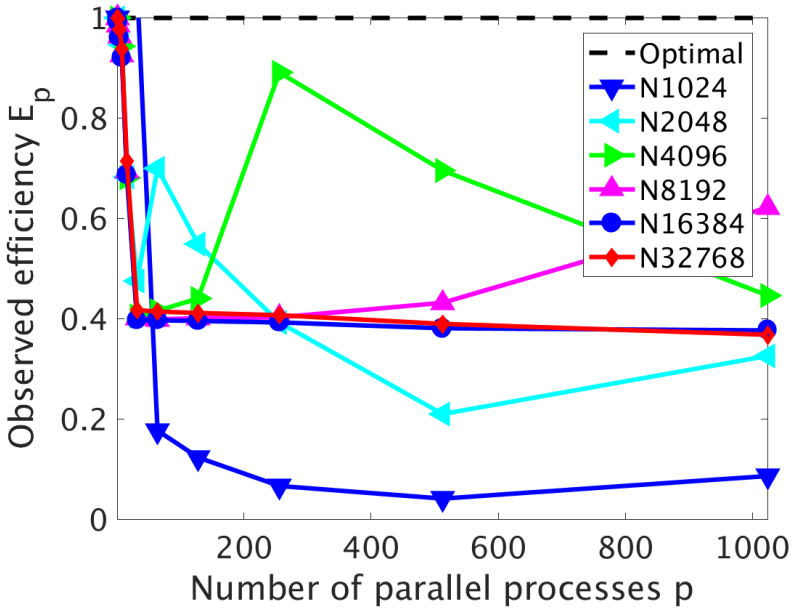
Speedup $S_p = T_1/T_p$											
N	1	2	4	8	16	32	64	128	256	512	1024
1024	1.00	2.81	5.25	10.14	19.17	33.92	11.31	15.75	16.96	21.00	88.20
2048	1.00	2.19	4.05	7.57	10.91	15.21	44.76	70.33	100.58	107.52	334.07
4096	1.00	2.03	4.01	7.55	10.90	12.97	26.57	56.38	228.21	356.33	456.43
8192	1.00	1.97	3.86	7.41	11.10	12.85	25.53	51.41	103.26	220.66	637.00
16384	1.00	1.98	3.84	7.37	10.99	12.70	25.43	50.65	100.51	194.73	385.80
32768	1.00	2.00	3.91	7.50	11.42	13.36	26.53	52.58	104.23	199.16	375.98

Efficiency $E_p = S_p/p$											
N	1	2	4	8	16	32	64	128	256	512	1024
1024	1.00	1.40	1.31	1.27	1.20	1.06	0.18	0.12	0.07	0.04	0.09
2048	1.00	1.10	1.01	0.95	0.68	0.48	0.70	0.55	0.39	0.21	0.33
4096	1.00	1.01	1.00	0.94	0.68	0.41	0.42	0.44	0.89	0.70	0.45
8192	1.00	0.99	0.96	0.93	0.69	0.40	0.40	0.40	0.40	0.43	0.62
16384	1.00	0.99	0.96	0.92	0.69	0.40	0.40	0.40	0.39	0.38	0.38
32768	1.00	1.00	0.98	0.94	0.71	0.42	0.41	0.41	0.41	0.39	0.37

taki 2018: strong scalability by p processes of MPI-only code



Speedup $S_p = T_1/T_p$



Efficiency $E_p = S_p/p$

Comparison of Performance Studies on Different Clusters

Poisson equation in 2-D, finite difference discretization, conjugate gradient method, matrix-free implementation, mesh resolution 4,096 × 4,096, system dimension 16,777,216 [HPCF–2008–1, HPCF–2010–2, HPCF–2015–6, HPCF–2018–10, HPCF–2018–18]

Cluster (year)	clock rate	serial (1 core) time	1 node all cores time (speedup)	32 nodes 1 core per node time (speedup)	32 nodes all cores time (speedup)
kali (2003)	(2.0 GHz)	02:00:49	N/A (N/A)	00:04:05 (29.59)	00:04:49 (25.08)
hpc (2008)		01:51:29	00:32:37 (3.42)	00:03:23 (32.95)	00:01:28 (76.01)
tara (2009)	(2.6 GHz)	00:31:16	00:06:39 (4.70)	00:01:05 (28.86)	00:00:09 (208.44)
maya 2009	(2.6 GHz)	00:17:05	00:05:55 (2.89)	00:00:35 (29.29)	00:00:08 (128.13)
maya 2010	(2.8 GHz)	00:17:00	00:05:48 (2.93)	00:00:34 (30.00)	00:00:06 (170.00)
maya 2013	(2.6 GHz)	00:12:26	00:02:44 (4.55)	00:00:15 (49.07)	00:00:12 (62.17)
taki 2018	(2.3 GHz)	00:13:32	00:01:03 (12.97)	00:00:14 (56.11)	00:00:02 (456.43)
Stampede2	(2.1 GHz)	00:13:04	00:01:00 (13.07)	00:00:16 (49.00)	00:00:01 (784.00)

- kali (2003): 2 processes per node, Myrinet interconnect
- hpc (2008): 4 processes per node, DDR InfiniBand interconnect
- tara (2009): 8 processes per node, QDR InfiniBand interconnect
- maya 2009: 8 processes per node, QDR InfiniBand interconnect
- maya 2010: 8 processes per node, DDR InfiniBand interconnect
- maya 2013: 16 processes per node, QDR InfiniBand interconnect
- taki 2018: 32 processes per node, EDR InfiniBand interconnect
- Stampede2 (2018): 32 processes per node, Intel Omni-Path (OPA) network

Conclusions

What does this mean for future research using taki?

- HPCF is providing three state-of-the-art clusters to the campus:
  a 173-node CPU cluster including 42 nodes with 36 cores per node,
  a GPU cluster including 4 Tesla V100 connected by NVLink, and
  an 8-node Big Data cluster with 48 TB disk space per node.
- *Weak scalability per node* holds, ideally with two or more MPI processes.
- *Strong scalability per node* is excellent up to approximately
  12 processes/threads per node = number of memory channels.
- *Strong scalability across multiple nodes* is excellent,
  demonstrating the quality of the EDR InfiniBand network.
- **Parallel code** benefits from HPCF2018, provided that the code takes advantage of the larger number of cores per node!
- **Serial code** benefits most from HPCF2013 due to higher clock rate.