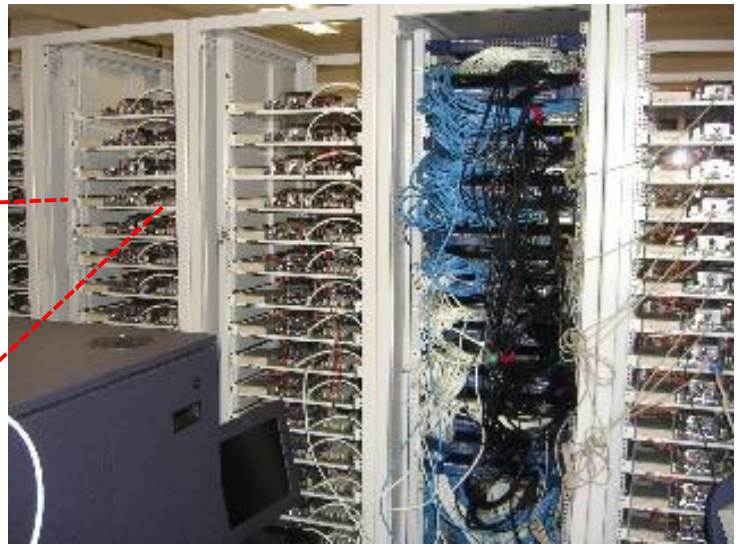
Trends of Network Topology on Supercomputers

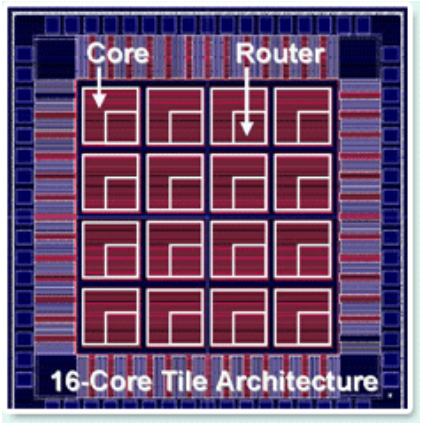
Michihiro Koibuchi
National Institute of Informatics, Japan
2018/11/27

From Graph Golf to Real Interconnection Networks

- Case 2: Supercomputer's Networks



- Case 1: On-chip Networks

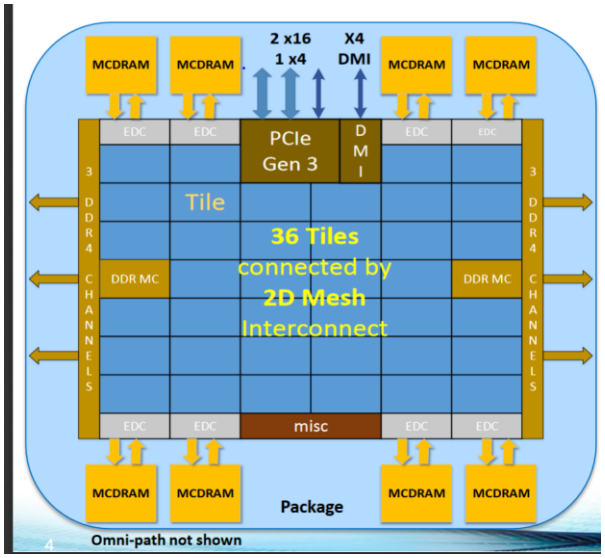


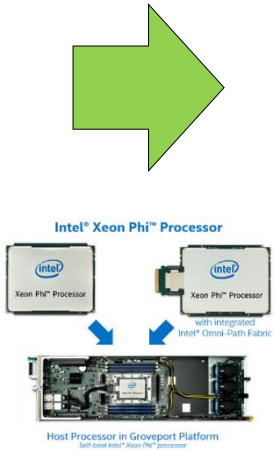
General Graph Category

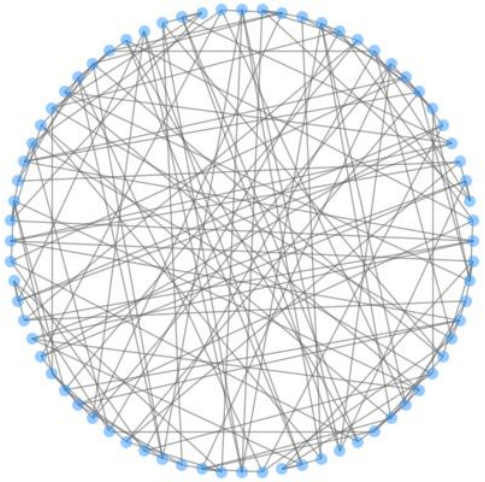
Case 1: (On-chip) Network Topology

- Intel Xeon Knights Landing (KNL)
  - 72 core, switch degree 4 → **72 vertices, 4 degrees**

1 title (vertices) = 2 cores





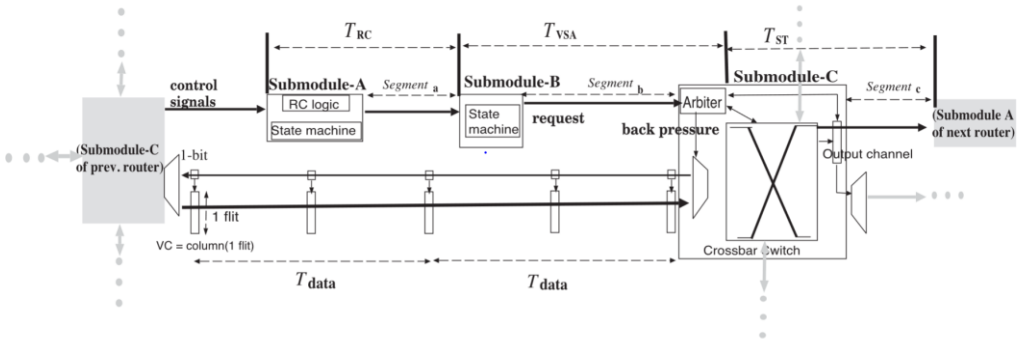


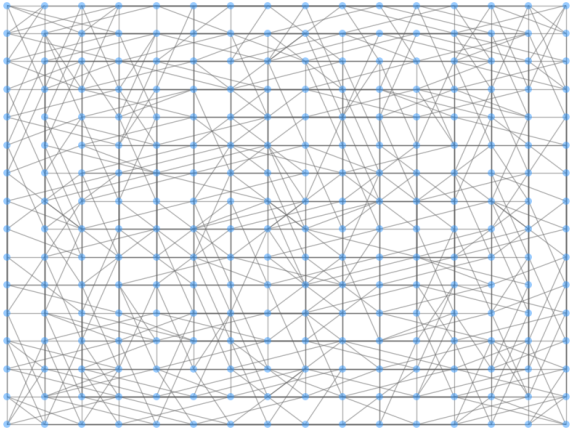
Discovery!
Dr. Nakao's Ideal Graph

https://www.alcf.anl.gov/files/HC27.25.710-Knights-Landing-Sodani-Intel.pdf

No More Regularity for real on-chip networks

- Moderate long cables
  - Enabled by recent technologies





- T. Krishna, et. al, Smart: Single-Cycle Multihop Traversals over a Shared Network on Chip, IEEE Micro 34(3): 43-56 (2014)
- R. Yasudo, et al, Scalable Networks-on-Chip with Elastic Links Demarcated by Decentralized Routers. IEEE Trans. Computers 66(4): 702-716 (2017)

- Routing Implementation
  - Source routing / Routing Table at each switch
  - When considering faulty node, irregularity will appear in regular topologies (fat tree, tori)

Case 2: Supecomputer's network
Cori – Cray XC40, Aries: 3019 switches, 30 degrees

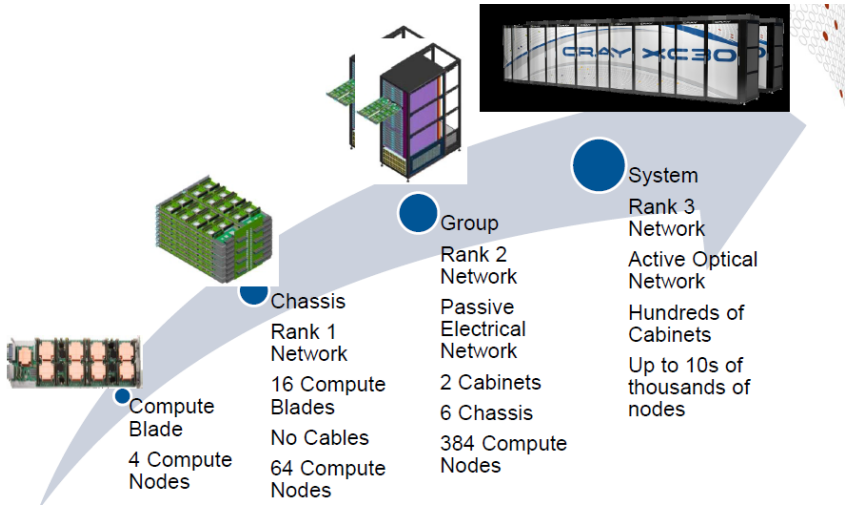
No.8 ranked in top500.org (Nov 2017)

Discovery!
Haruishi's graph minimizes network diameter (3)

Dragonfly + Aries interconnect.
- 1 compute blades = 1 aries switch = 4 compute nodes
- 1 cabinet = 3 chassis = 48 compute blades

# of nodes/switches
- 1 Haswell node = 2 Xeon™ Processor E5-2698 v3 (16-core)
- 1 KLN node = Intel® Xeon Phi™ Processor 7250 (68 cores)
- 14 Haswell cabinets → 2388 compute node[1]
- 54 KLN cabinets → 9688 compute node [1]

Degree 30
- 6 aries with fully connected: 15 links/switch
- 6 chassis with copper cables: 5 links/switch
- Active optical cables interconnect groups:10 links/switch
- Degree 30 but each switch has 48 ports (for 12 cores?)



[1] http://www.nersc.gov/users/computational-systems/cori/configuration/

Cray XC40*, Aries interconnect (Dragonfly topology)

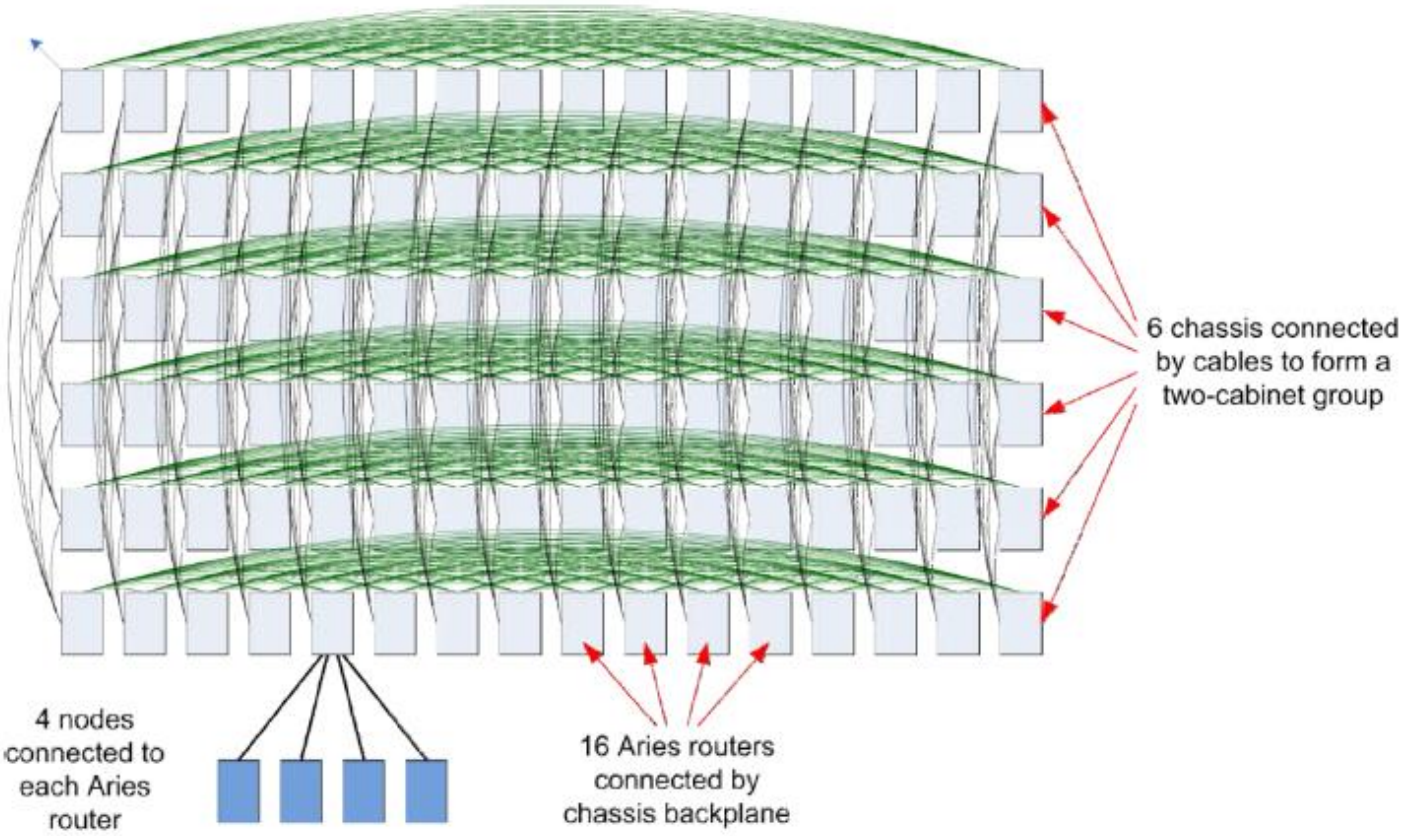


Figure 8: Structure of a Cray XC system's electrical group. Each row represents the 16 Aries in a chassis with four nodes attached to each and connected by the chassis backplane. Each column represents a blade in one of the six chassis per two-cabinet group, connected by copper cables.

https://www.cray.com/sites/default/files/resources/CrayXCNetwork.pdf

The other top-10 supercomputers

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National S China	10,649,600	93,014.6	125,435.9	15,371
2	**Tianhe-2** 12C 2.200 National Super Com China				
3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272
4	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , ExaScaler Japan Agency for Marine-Earth Science and Technology Japan	19,860,000	19,135.8	28,192.0	1,350
5	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , Cray Inc. DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209
6	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , IBM DOE/NNSA/LLNL United States	1,572,864	17,173.2	20,132.7	7,890
7	**Trinity** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/NNSA/LANL/SNL United States	979,968	14,137.3	43,902.6	3,844
8	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/SC/LBNL/NERSC United States	622,336	14,014.7	27,880.7	3,939
9	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path , Fujitsu Joint Center for Advanced High Performance Computing Japan	556,104	13,554.6	24,913.5	2,719
10	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect , Fujitsu RIKEN Advanced Institute for Computational Science (AICS) Japan	705,024	10,510.0	11,280.4	12,660

GraphGolf 2018 problem

40,256,4-radix switches? a bit complex for graphgolf

GraphGolf2017: (4896,24) (Order,degree)

The detail of the network is not opened

GraphGolf2017: (9344,10)

GraphGolf2017: (98304,10)

384-pt sw + 48-pt sw: a bit complex for graphgolf

GraphGolf2017: (88128,12)

https://www.top500.org/lists/2017/11/

Rank 3. Piz Daint – Cray XC40/XC50, Aries

1726 switch, degree 30

Ref [1] Dropbox (KoibuchiLab)/NII/Docs/2016/1226-graphgolf-survey/Nguyen-survey/Cray_XCNetwork.pdf
[2] https://www.cscs.ch/computers/piz-daint/
[3] https://www.cscs.ch/computers/dismissed/piz-daint-piz-dora/
[4] cray-xc50-Product-Brief.pdf

Dragonfly + Aries interconnect.

- 1 compute blades = 1 aries switch = 4 compute nodes
- 1 cabinet = 3 chassis (16 blade/chasis, 1blade=1aries) = 48 compute blades

# of nodes/switches?

- 1 XC50 node = Xeon E5-2690v3 CPU (12 cores) + NVIDIA® Tesla® P100 [2]
  - Past: Xeon® E5-2670 8 core+ NVIDIA® Tesla® K20X [3]
  - 5320 compute nodes [2] ~ 28 cabinet = 1344 switches
- 1 XC40 node = 2 Intel® Xeon® E5-2695 (36 cores)
  - 1431 compute nodes [2] ~ 8 cabinet = 382 switches.

Degree? 30

- 6 aries with fully connected: 15 links/switch
- 6 chassis with copper cables: 5 links/switch
- Active optical cables interconnect groups:10 links/switch
  - 10 global links
- Degree 30 but each switch has 48 ports [4] (for 12 cores?)

Upgrade history [2]
+ November 2013: Upgrade to a hybrid architecture and extension from 12 to 28 cabinets.
+ November 2016: Update of processors (CPU) and accelerators (GPU) and combination with the Piz Dora supercomputer > Cray XC40/XC50.

Network
The system has Aries routing and communications ASIC, with Dragonfly network topology.

Rank 7 Trinity system: 4855 switches, degree 30

Ref [1] Dropbox (KoibuchiLab)/NII/Docs/2016/1226-graphgolf-survey/Nguyen-survey/Cray_XCNetwork.pdf
[2] http://www.lanl.gov/projects/trinity/specifications.php
[3] http://www.lanl.gov/projects/trinity/_assets/docs/trinity-overview-for-web.pdf

Dragonfly + Aries interconnect.

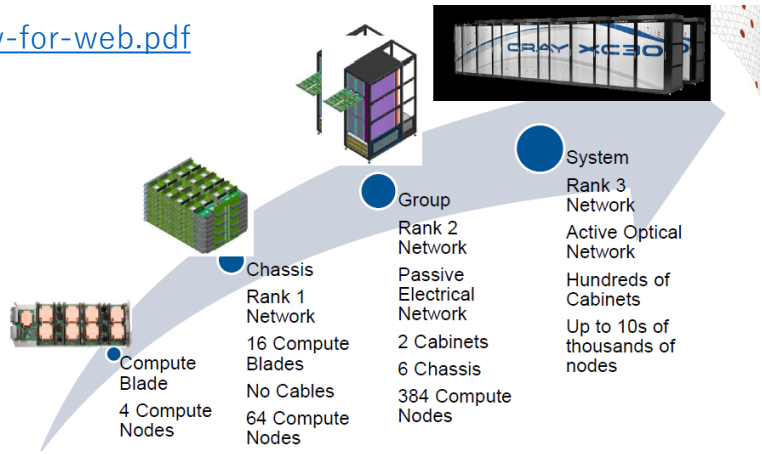
- 1 compute blades = 1 aries switch = 4 compute nodes
- 1 cabinet = 3 chassis = 48 compute blades

# of nodes/switches

- 1 KLN node = Intel® Xeon Phi™ Processor 7250 (68 cores)
  - Pass: 1 Haswell node = 2 Xeon™ Processor E5-2698 v3 (16-core)
- 19,420 nodes [2]

Degree? 30

- 6 aries with fully connected: 15 links/switch
- 6 chassis with copper cables: 5 links/switch
- Active optical cables interconnect groups:10 links/switch
- Degree 30 but each switch has 48 ports



Metric	Trinity	Haswell Partition	KNL Partition
Node Architecture	KNL + Haswell		
Memory Capacity	2.11 PB	> 1 PB	>1 PB
Memory BW	>6 PB/sec	> 1 PB/s	>1PB/s + >4PB/s
Peak FLOPS	42.2 PF	11.5 PF	30.7 PF
Number of Nodes	19,000+	>9,500	>9,500
Number of Cores	>760,000	>190,000	>570,000

Rank. 9 Oakforest-PACS: 12 384-port sw +342 12-port sw

→ Not applicable for graphgolf

Ref [1] https://www.ccs.tsukuba.ac.jp/wp-content/uploads/sites/14/2016/05/boku.pdf

Total peak performance			25 PFLOPS
Total number of compute nodes			8,208
Compute node	Product		Fujitsu Next-generation PRIMERGY server for HPC (under development)
	Processor		Next-generation of Intel® Xeon Phi™ (Code name: Knights Landing), >60 cores
	Memory	High BW	16 GB, > 400 GB/sec (MCDRAM, effective rate)
		Low BW	96 GB, 115.2 GB/sec (DDR4-2400 x 6ch, peak rate)

Fat-tree with (completely) full-bisection bandwidth

1 compute node = = Intel® Xeon Phi™ Processor 7250 (68 cores)

# of nodes/switches?

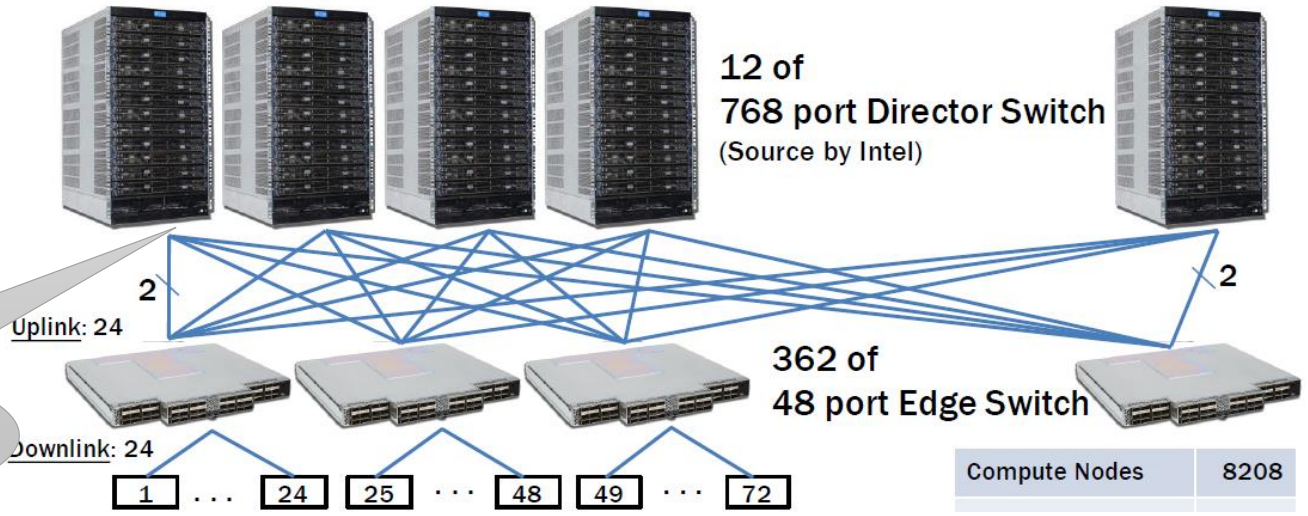
- 8208 nodes
- 362 switches (48 port)

Degree? 48

- 24 for nodes
- 24 for network

Diameter: 2 or 4

2 cables per link (Link Aggregation)



→ Kohta Nakashima(FujitsuLab.), CSA Keynote (15:15–16:00, 29 Nov)

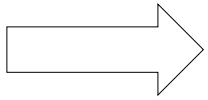
Summary for General Graph Category

Order/degree pairs (n, d) in GraphGolf 2018

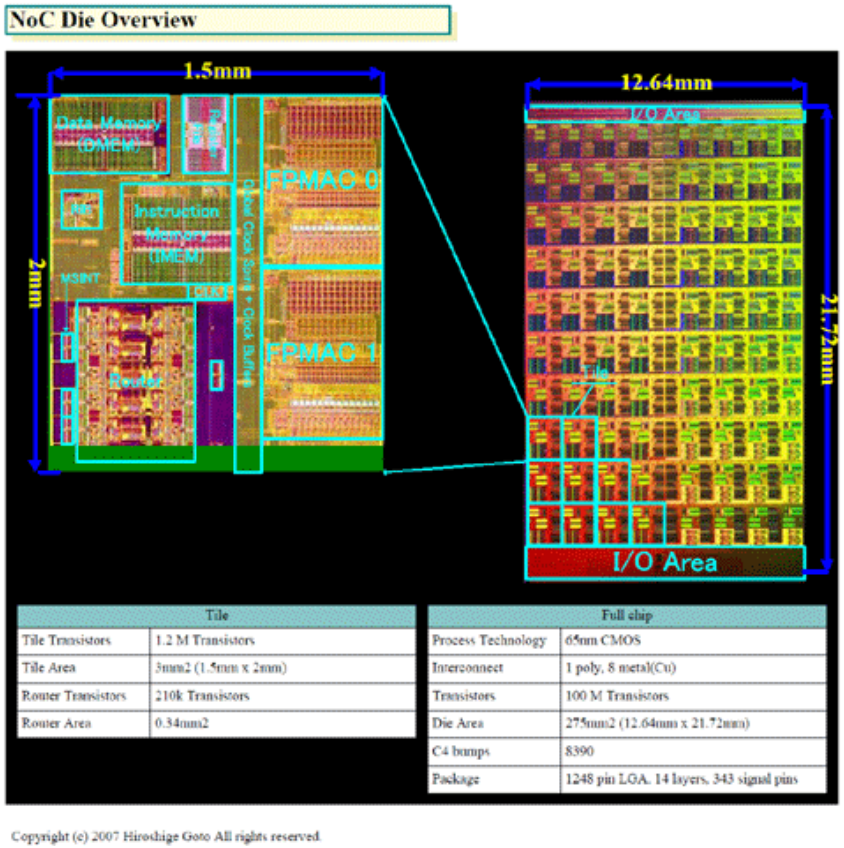
- (72, 4) On-chip network on Xeon KNL processor
- (3019, 30), (4855, 30) Top10 supercomputers
- (200000, 32), (200000, 64), (400000, 32) Exascale Supercomputers (to appear in 2020th)

Grid Graph Category

Intel 80-tile chip
10x8 2D Mesh
Chip is Rectangle!!



Order/degree/length pairs (w × h, d, r) =
(4 × 16, 4, 4),
(32 × 32, 4, 3/4/5),
(16 × 64, 4, 4/5/7),
(4 × 256, 4, 12/18/24)



https://pc.watch.impress.co.jp/docs/2007/0215/kaigai337_01.pdf

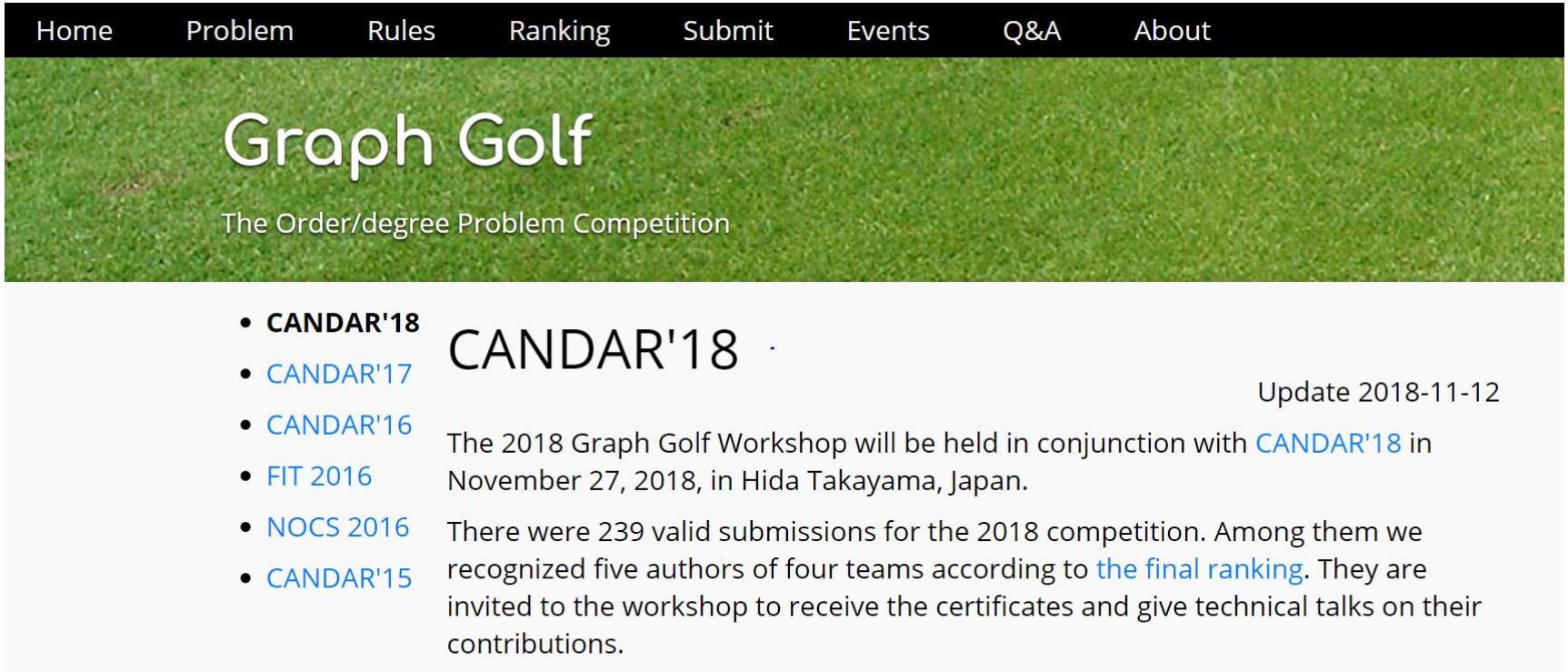
Grid Graph Category's lower bound*1

no.	Grid Size	degree	length	BOTH	MOORE	DIST	comment
0	4x16	4	4	3.063492	2.857143	2.5625	balanced
1	32x32	4	3	7.68564	5.300098	7.444892	DIST dominant
1	32x32	4	4	6.289643	5.300098	5.708944	balanced
1	32x32	4	5	5.681585	5.300098	4.667495	MOORE dominant
2	16x16	4	4	7.492932	5.300098	7.042278	DIST dominant
2	16x16	4	5	6.498805	5.300098	5.734127	balanced
2	16x16	4	7	5.635291	5.300098	4.239297	MOORE dominant
3	4x256	4	12	8.306379	5.300098	7.685427	DIST dominant
3	4x256	4	18	6.42249	5.300098	5.295615	Balanced
3	4x256	4	24	5.708368	5.300098	4.102891	MOORE dominant

*1 K. Nakano et. al, Randomly Optimized Grid Graph for Low-Latency Interconnection Networks. ICPP 2016: 340-349

Summary

- Graph golf is filling the gap between theory and real products of supercomputers
- All the materials are available from
  - http://research.nii.ac.jp/graphgolf/events.html



https://www.nii.ac.jp/news/release/2018/1127.html

NEWS RELEASE

NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

2018 年（平成 30 年）11 月 27 日

次世代のスパコン設計を模した 40 万頂点数の巨大グラフを発見

通信遅延の大幅な低下などの実用に期待

～効率的なスパコン設計につながるグラフ発見を競うコンペ「グラフ ゴルフ」で～

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所（NII、所長：喜連川 優、東京都千代田区）は、複雑なネットワーク構成をスイッチ間の接続関係を表す簡単なグラフ(*1)に抽象化し、より単純な構成のグラフの発見を競うコンペティション「グラフ ゴルフ」(*2) で優れたグラフを発見した 3 名の個人と 1 チームを、本日 11 月 27 日、岐阜県高山市で開催された国際シンポジウム「CANDAR2018」(*3) で表彰しました。これらのグラフは、効率的なプロセッサーコア間の通信や、スーパーコンピューター（スパコン）の超並列計算の最長通信時間の最小化、次世代のスパコン設計の通信遅延の低下など、性能向上への応用が期待されます。また今回、表彰者の一人である北須賀 輝明が発見した 4 つのグラフが、グラフ理論分野において著名な問題とされてきた次数直径問題の最大グラフ(*4) の記録更新になるという新たな展開が生まれ、グラフに関する理論分野にも貢献しています。

最近のコンピューターは大規模で複雑になってきており、特にスパコンでは 1 千万以上のプロセッサーコア（以降、コアと表記）が接続されるものも登場しています。しかし、一つのコアに直接接続できるコアには制限があることから、コア間のネットワーク構成を工夫して膨大な数のコアを効率的に相互接続することが、スパコンの処理能力に大きく影響します。本コンペでは、スパコンの専門家でなくて

Call for Papers

http://www.ieice.org/eng/s_issue/cfp/2019_12ED.pdf

—— Special Section on Parallel and Distributed Computing and Networking ——

The IEICE Transactions on Information and Systems announces that it will publish a special section entitled "Special Section on Parallel and Distributed Computing and Networking" in December 2019.

The IEICE Transactions on Information and Systems announces a forthcoming Special Section on "Parallel and Distributed Computing and Networking" to be published in December 2019. The objective of this special section is to publish and overview recent progress in the interdisciplinary area of Parallel and Distributed Computing and Networking. This special section will include papers based on the presentation at the International Symposium on Computing and Networking (CANDAR'18) and International Conference on Field-Programmable Technology (FPT'18) in addition to papers applied for this call for papers. All submitted papers are subjected to the same review process as those papers accepted for publication in the regular issues.

1. Scope

This special section aims at timely dissemination of research in these areas. Possible topics include, but are not limited to:

a. Parallel/distributed Algorithms and Applications (e.g., High-performance computing, Image processing and Computer graphics, Data mining and Information retrieval, Multicore and Accelerator-based computing, Network algorithms, Green computing, Simulation and Visualization, Scheduling and Load balancing, and Performance model and Evaluation),
b. Parallel/distributed Systems and Architectures (e.g., Parallel processor architectures, Cluster/Grid systems, Network and Storage architectures, Network-on-Chip, and High performance interconnect, Reconfigurable system),
c. Distributed Systems and Networking (e.g., Ubiquitous computing, P2P networks, and Wireless networks and Mobile computing),
d. Software and Technologies for Parallel/Distributed Systems (e.g., Operating systems, Middleware, Tools, Virtualization, Parallel programming models and Languages, Web services, Cloud and Distributed computing, and Cluster/Grid Scheduling and Resource management).

2. Submission Instructions

– A manuscript should be prepared according to the guideline given in "The Information for Authors" (http://www.ieice.org/eng/shiori/mokuji_iss.html). We encourage the authors to use the IEICE Style File (http://www.ieice.org/ftp/index-e.html). The preferred length of the manuscript is 8 pages for a PAPER and 2 pages for a LETTER with the format determined by the IEICE Style File.
– Submit the manuscript through the IEICE Web site (https://review.ieice.org/regist/regist_baseinfo_e.aspx). Choose ""[Special-PA] Parallel and Distributed Computing and Networking" in the menu of "Journal/Section" in the submission page. Do not choose "[Regular-ED] Information and Systems" or other special sections.
–Authors must agree to the "Copyright Transfer and Page Charge Agreement" via electronic submission.
–Submission deadline of the manuscript is January 7, 2019 (No Extension).

Contact:
Michihiro Koibuchi, Information Systems Architecture Research Division,
Tel: +81-3-4212-2575, Fax: +81-3-4212-2035, Email: pdcn2019@nii.ac.jp

3. Special Section Editorial Committee

Guest Editor-in-Chief: Michihiro Koibuchi (National Institute of Informatics)
Guest Editor: Fukuhito Ooshita (Nara Institute of Science and Technology), Shinya Takamaeda (Hokkaido University)
Guest Associate Editors:
Ikki Fujiwara (NII), Naoto Fukumoto (FUJITSU Lab.), Shugo Ogawa (NEC), Hiroshi Inoue (IBM Japan),
Jun Kawahara (Nara Institute of Science and Technology), Hiroshi Yamamoto (Ritsumeikan University),
Yasuaki Ito (Hiroshima Univ.), Kenji Kise (Tokyo Tech), Teruaki Kitasuka (Hiroshima Univ.),
Hironori Nakajo (Tokyo Univ. of Agriculture and Technology), Yoshitaka Nakamura (Future Univ. Hakodate),
Takashi Miyoshi (FUJITSU Lab.), Yukinori Sato (Toyohashi University of Technology), Ryota Shioya (Univ. of Tokyo),