Camphor2：OFPと同じXeon Phi KNLを搭載し，OFPより少し長く運用される京都大学スパコン

深沢 圭一郎

京都大学学術情報メディアセンター

Camphor2

Xeon Phi KNLのスパコン

京大スパコンでは、2014年に補正予算でXeon Phi KNCを導入し、2016年にリプレイスで現行サブシステムAであるXeon Phi KNLを導入しました。
→このサブシステムAがCamphor2という名前です。

KNCと異なり、KNLはコプロセッサではなく、x86コードがそのまま動作するという魅力がありました。

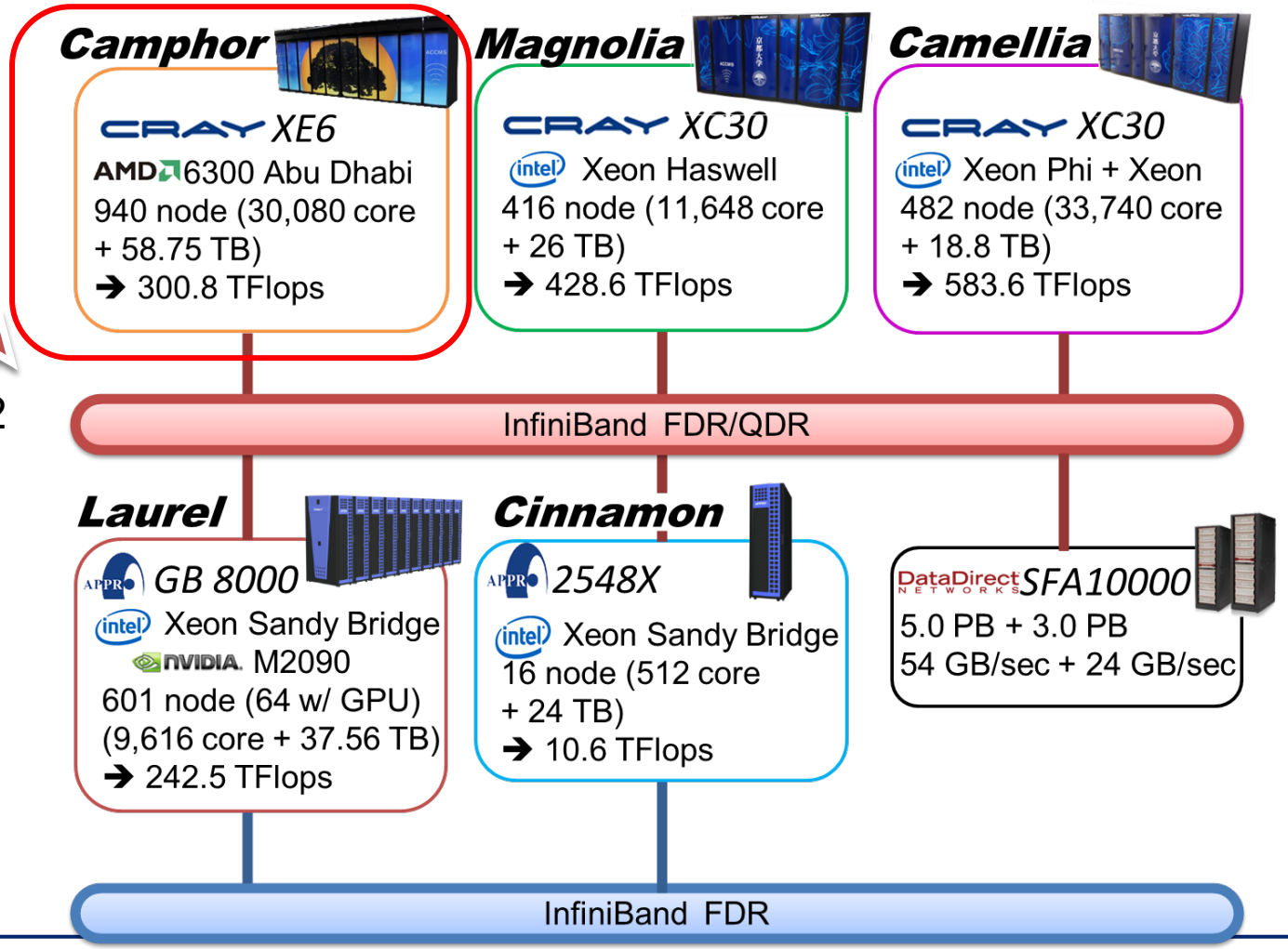
仕様書では、Xeon Phi KNLを想定しつつ、少しインターコネクトの性能を高く要求した結果、Cray XC40が導入されました。

Camphor2の筐体

Xeon Phi KNL導入前

2012~2016に稼働していたシステム

この後継機がCamphor2



Xeon Phi KNL導入後

2016~2022に稼働しているシステム

Camphor 2 (System A)

CRAY *XC40*

intel Xeon Phi KNL 68cores 1.4GHz x 1 /node

#nodes	= 1,800
#total cores	= 68 cores x 1,800 → 122,400 cores
Peak performance	= 3.05TFops x 1,800 → 5.48 PFlops
Memory capacity	= (96+16 GB) x 1,800 → 196.9 TB
Burst buffer	= 230 TB, 200 GB/sec DATAWARP

Storage

DataDirect NETWORKS *ExaScaler (SFA14K)*

Disk capacity	= 24 PB
Bandwidth	= 150 GB/sec
Burst buffer	= 230 TB, 250 GB/sec

IME

高速通信網 InfiniBand EDR/FDR

高速通信網 Omni-Path

Laurel 2 (System B)

CRAY *CS400 2820XT*

intel Xeon Broadwell 18cores 2.1GHz x 2 /node

#nodes	= 850
#total cores	= 36 cores x 850 → 30,600 cores
Peak performance	= 1.21 TFlops x 850 → 1.03 PFlops
Memory capacity	= 128 GB x 850 → 106.3 TB

Cinnamon 2 (System C)

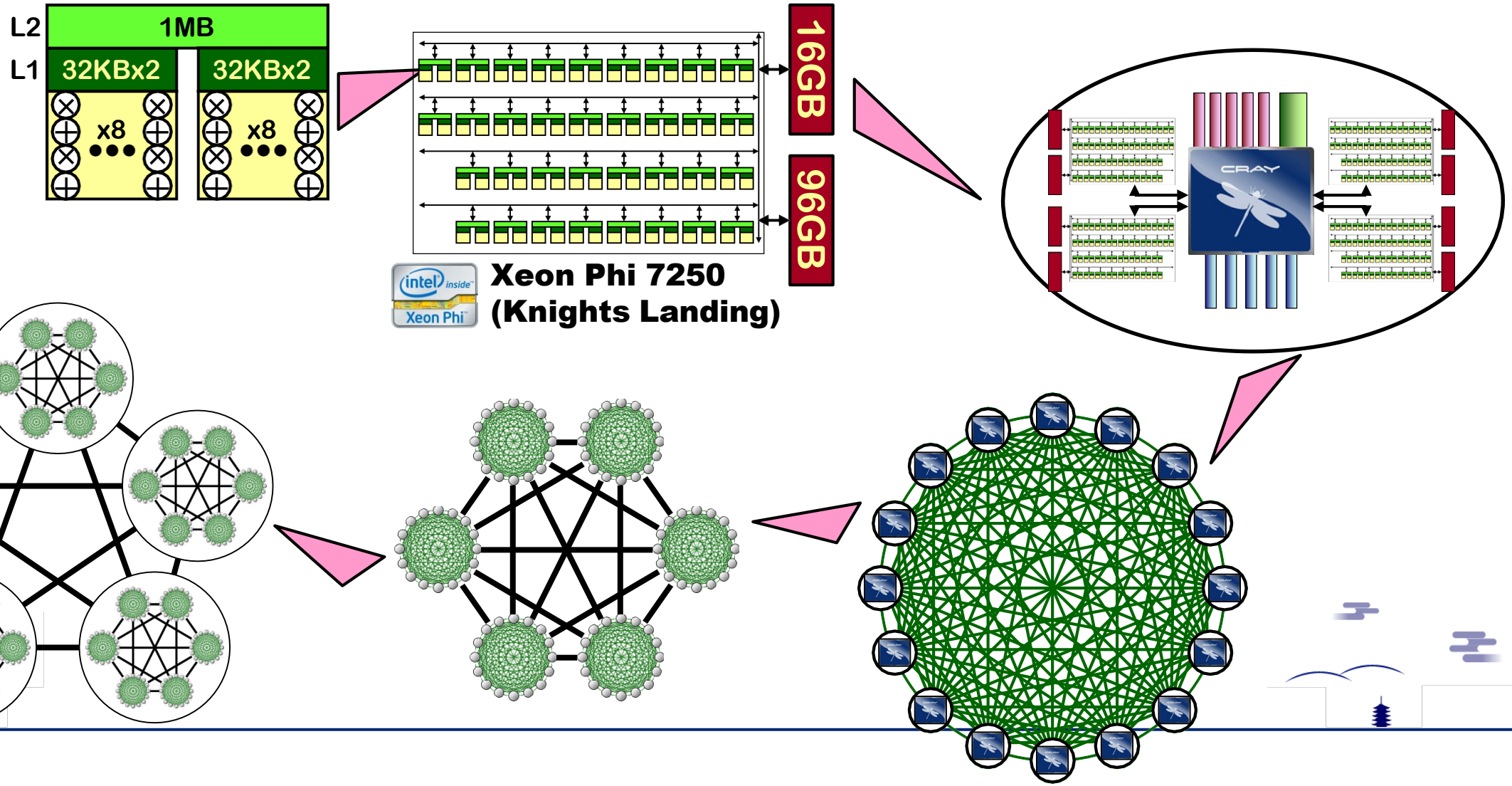
CRAY *CS400 4840X*

intel Xeon Haswell 18cores 2.3GHz x 4 /node

#nodes	= 16
#total cores	= 72 cores x 16 → 1,152 cores
Peak performance	= 2.65 TFlops x 16 → 42.4 TFlops
Memory capacity	= 3 TB x 16 → 48.0 TB

Camphor2のOFPとの違い

インターコネクトがCray独自のAries+Dragonflyトポロジー



サブシステムA－Camphor2の立ち位置

自前のコードを動かすメニーコアCPUスパコン

ISVアプリや汎用的に利用されているアプリケーションは置いておいて、自作のシミュレーションコードや計算コードなどが早く走ることを目標にしたシステム。
→汎用機はサブシステムB（Xeonクラスタ）

そのため、高Flops値、高メモリバンド幅のXeon Phi KNLが最適でした。
→Camphorから10倍以上のFlops値向上

ただ、Opteron Abu Dhabiとは構成が異なるため、最適化をがんばらないと性能が出ませんでした。

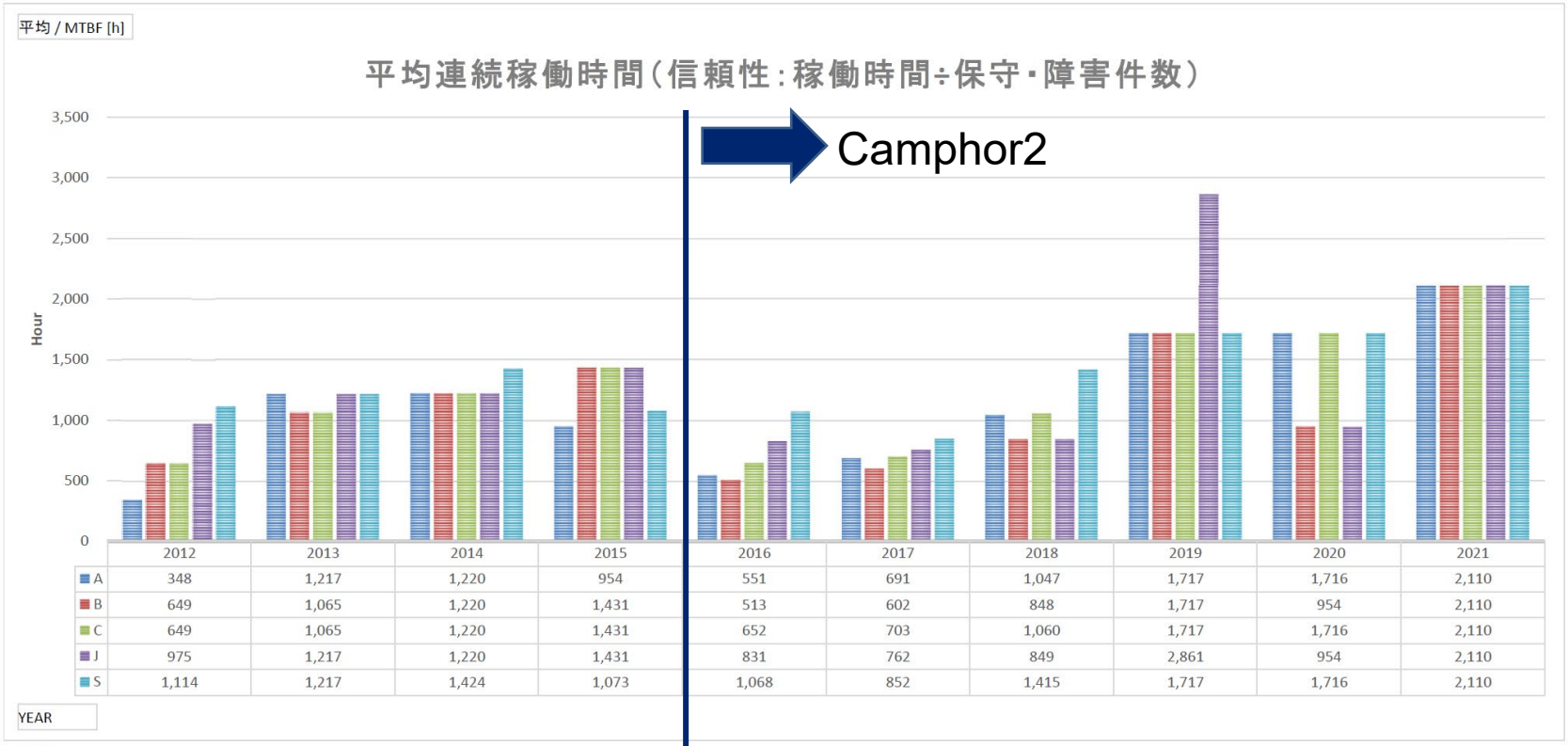
Camphor2の性能比較

MHDコードを用いた様々な計算機システム上での性能比較

	Core/CPU	Rpeak [TFlops]	Rmax [TFlops]	Rmax /CPU [GFlops]	Eff. [%]	B/F	Suitable domain decomposition	CPU architecture
SX-ACE	1024/256	65.50	29.20	114.0	45	1.00	3D xyzm	Vector
SX-AT	1024/128	314.37	61.47	480.2	20	0.50	2D xyzm	Vector 20B
K	262144/32768	4194.30	914.12	27.9	22	0.50	3D mxyz	SPARC64 VIIIfx
FX100	16384/512	576.72	91.49	178.7	17	0.42	3D xyzm	SPARC64 XIfx
EPYC7282	32/2	1.43	0.20	99.9	14	0.12	3D xyzm	Approx. AVX2
XC30	448/32	16.49	1.37	42.8	8	0.11	2D xyzm	Xeon (Haswell)
ITO-A	72000/4000	6912.00	470.10	117.5	7	0.07	1D xyzm	Xeon (Skylake)
XC40	1088/16	48.86	4.32	**273.3**	9	0.16	3D xyzm	Xeon Phi KNL
Tesla K20X	2688/1	1.31	0.15	153.3	12	0.19	3D xyzm	Kepler
ITO-B GPU	3584/1	5.30	0.38	382.2	7	0.14	3D xyzm	Pascal
DGX-A100	27648/8	78.08	8.90	1113.0	11	0.16	3D xyzm	Ampere
ThunderX2	256/8	0.70	4.50	86.9	16	0.30	3D mxyz	Arm v8
FX700	192/4	1.70	11.06	425.5	15	0.37	3D xymz	A64FX

Camphor2（+その他）の稼働状況

2017年度には、全系が止まる障害が65時間もあり、当初は安定しないシステムでした（細かい障害はありまくり）。



	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
A	348	1,217	1,220	954	551	691	1,047	1,717	1,716	2,110
B	649	1,065	1,220	1,431	513	602	848	1,717	954	2,110
C	649	1,065	1,220	1,431	652	703	1,060	1,717	1,716	2,110
J	975	1,217	1,220	1,431	831	762	849	2,861	954	2,110
S	1,114	1,217	1,424	1,073	1,068	852	1,415	1,717	1,716	2,110

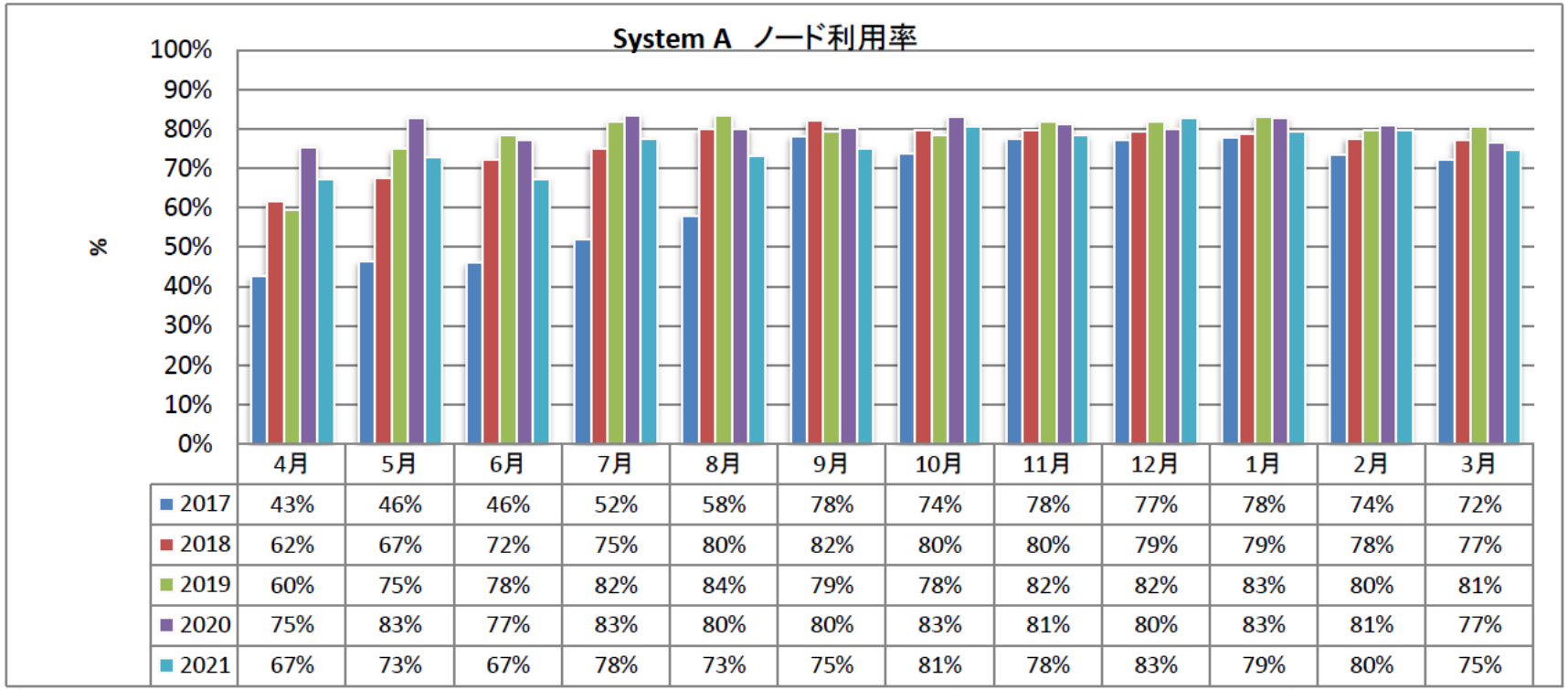
参考

期間	時間換算	
1週間	168	時間
2週間	336	時間
3週間	504	時間
4週間	672	時間
5週間	840	時間

京都大学
KYOTO UNIVERSITY

Camphor2の利用状況

段々と使われるようになりました



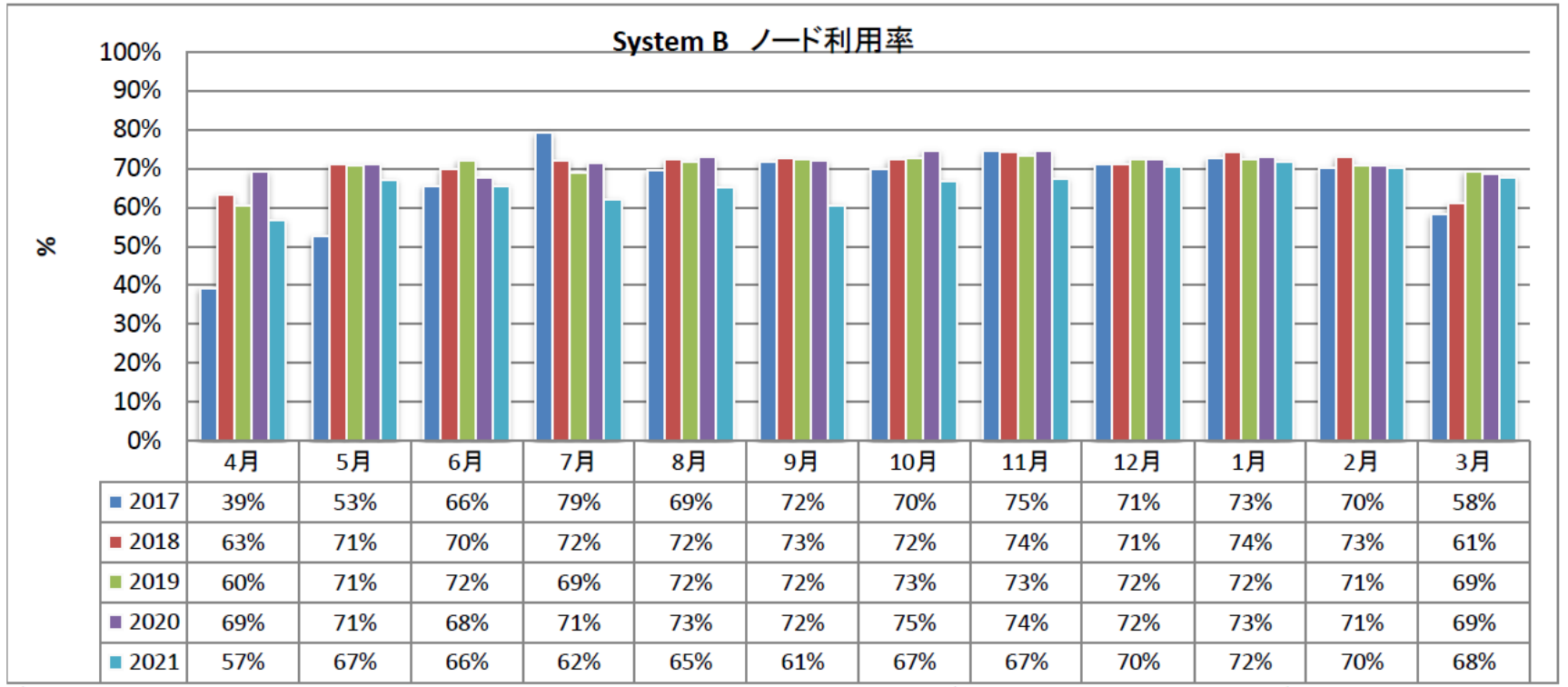
	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月
2017	43%	46%	46%	52%	58%	78%	74%	78%	77%	78%	74%	72%
2018	62%	67%	72%	75%	80%	82%	80%	80%	79%	79%	78%	77%
2019	60%	75%	78%	82%	84%	79%	78%	82%	82%	83%	80%	81%
2020	75%	83%	77%	83%	80%	80%	83%	81%	80%	83%	81%	77%
2021	67%	73%	67%	78%	73%	75%	81%	78%	83%	79%	80%	75%

※ノード利用率：バッチジョブの利用ノード数/バッチジョブ対象ノード数

年間平均

	2015	2016	2017	2018	2019	2020	2021
Node	61%	54%	65%	76%	79%	80%	76%

（参考情報）Laurel2の利用状況



	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月
2017	39%	53%	66%	79%	69%	72%	70%	75%	71%	73%	70%	58%
2018	63%	71%	70%	72%	72%	73%	72%	74%	71%	74%	73%	61%
2019	60%	71%	72%	69%	72%	72%	73%	73%	72%	72%	71%	69%
2020	69%	71%	68%	71%	73%	72%	75%	74%	72%	73%	71%	69%
2021	57%	67%	66%	62%	65%	61%	67%	67%	70%	72%	70%	68%

年間平均

※ノード利用率：バッチジョブの利用ノード数/バッチジョブ対象ノード数

	2015	2016	2017	2018	2019	2020	2021
Node	64%	53%	66%	71%	71%	72%	66%

JHPCNやHPCIへの資源提供

高い演算性能のため、HPCIへの提供資源としました。

JHPCN（2021年度）

- 128 ノード（8,704コア、390.4TFLOPS）×1年
- 128 ノード（8,704コア、390.4TFLOPS）×8 週

HPCI （2021年度）

- 720ノード (48,960コア、2,193TFLOPS)×1年

それぞれ専用キューを用意して、
課題間で共有の提供（通常運用とは別枠）

年度別利用件数

年度	JHPCN	HPCI
2017	2	4
2018	4	4
2019	4	5
2020	4	8
2021	7	4
2022	5	2

まとめ

もうすぐ運用を終える京大のXeon PhiスパコンCamphor2

- ✓ 2016年当時で3TFlopsの高いCPU性能
  →今が高すぎて、次期システムの性能アップ率が高くならない…
- ✓ 東大や北大と異なり、Cray独自のインターコネクトを実装。
- ✓ ピーキーなマシンだったので、自前コード向けでスパコンぽかった。
- ✓ 動作が安定しないなどありましたが、最近はこなれて、よく利用されています。
- ✓ 今年度もうすぐ停止し、新しいCamphor3?になります。
- ✓ 見学はお早めに!