2016年12月16日
第16回PCクラスタシンポジウム@ 秋葉原コンベンションホール

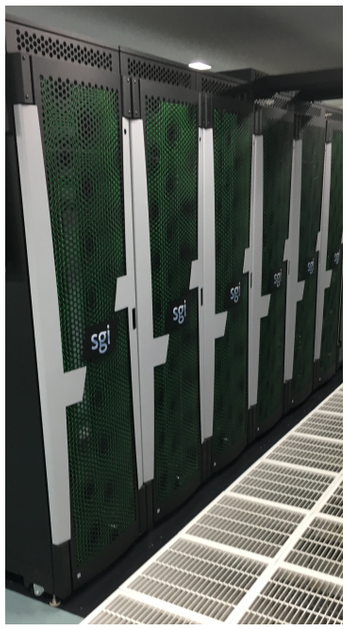
実用アプリケーション・クラウド採択課題
「各種クラウドサービス・FOCUS・Oakleaf-FX10での OpenFOAM 性能・費用ベンチマークテスト」

オープンCAE学会V&V委員会
今野　雅（株式会社OCAEL，東京大学客員研究員）

Copyright © 2016 OCAEL All rights reserved.

研究の背景

- **大学のスパコン：東京工業大学TSUBAME，東京大学FX10・Reedbush等**
  - ✓ 産業利用など教育・公共機関以外でも利用可能
  - ✓ 通常，課題審査が必要
  - ✓ 通常，使った分だけ課金ではなく，1ヶ月～1年単位での課金
- **産業界専用の公的スパコン：FOCUS**
  - ✓ 法人の場合，課題の審査無く利用可能
- **クラウドサービス：Amazon EC2，Microsoft Azure等**
  - ✓ 手持ちの計算機リソースでは難しい中・大規模な解析に適する
- 各スパコン，クラウドではCPU性能やインターコネクトに違いがあり，利用料金も当然異なる



オープンCAE学会で共通OpenFOAMベンチマークを作成し比較した

Copyright © 2016 OCAEL All rights reserved.

対象システムの特徴

- **東京大学 Reedbush-U，Oakleaf-FX**
  - ✓ 利用には課題申請と審査が必要.
  - ✓ FXはCPUコア単体の性能が低いので，非並列の前処理・後処理が遅い
- **東京工業大学 TSUBAME**
  - ✓ 学術利用以外では課題申請が必要.
  - ✓ 様々なシステムとキューがあり自由度が高いが課金体系は多少複雑
- **FOCUS**
  - ✓ 使った分だけ1円単位で課金.
  - ✓ 法人のみ使用可能．課題申請は不要.
- **Amazon EC2 (クラウド)**
  - ✓ 使った分だけ1時間単位で課金
  - ✓ 入札で価格が決まる安価なスポット利用有り．変動が激しいので今回は除外.
- **Microsoft Azure (クラウド)**
  - ✓ 使った分だけ分単位で課金
  - ✓ 高速なインターコネクトを持つHPC向けインスタンス有り(今回検討したA9)

Copyright © 2016 OCAEL All rights reserved.

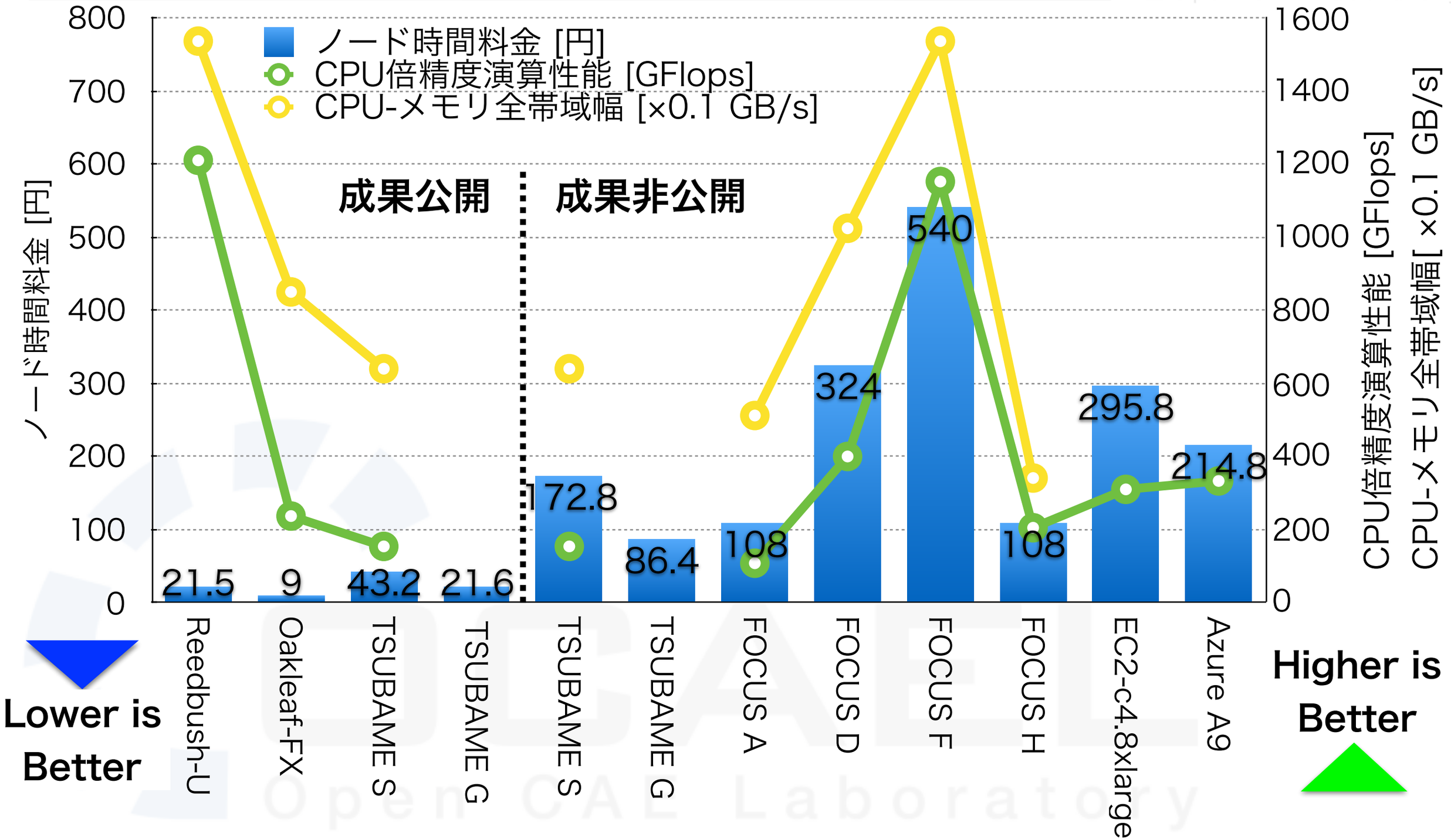
対象システムのノード・インターコネクト性能

機関	システム	CPU [GPU] (周波数[GHz])	CPU数 (コア)	倍精度性能 [GFlops]	メモリ[GiB] (帯域幅[GB/s])	インターコネクト (帯域幅[Gbps])
東京大学	Reedbush-U	Intel Xeon E5-2695 v4 (2.1-3.3(※1))	2(36)	1210	256 (76.8×2)	Infiniband EDR(100)
	Oakleaf-FX	Fujitsu SPARC64 IXfx (1.848)	1(16)	237	32 (85)	Tofu(40)×双方向 ×10(4方向同時通信)
東京工業大学	TSUBAME S	Intel Xeon E5-2670 (2.93-3.2(※1))	2(12)	154	54 (32×2)	Infiniband QDR (40)×2
	TSUBAME G	[nVIDIA Tesla K20X] (0.732)	3GPU	1310×3	6×3 (150×3)	
FOCUS	A	Xeon L5640(2.26)	2(12)	108	48 (25.6×2)	Infiniband QDR(40)
	D	E5-2670 v2(2.5)	2(20)	400	64 (51.2×2)	Infiniband FDR(56)
	F	E5-2698 v4(2.2)	2(40)	1152	128 (76.8×2)	
	H	D-154(2.1)	1(8)	205	64 (34.1)	10GbE(10)×2 or 4
Amazon	EC2 c4.8xlarge	Intel Xeon E5-2666 v3(2.9)(※2)	2(18)	310(※3)	60 (不明)	10GbE(10)
Microsoft	Azure A9	Intel Xeon E5-2670(2.6) (※2)	2(16)	333	112 (不明)	Infiniband QDR(40)

(※1)ターボブースト(※2)仮想マシン．Hyper-threading無効(※3)Intel MKL LINPACK測定値

Copyright © 2016 OCAEL All rights reserved.

ノード時間料金・CPU演算性能・メモリ帯域幅

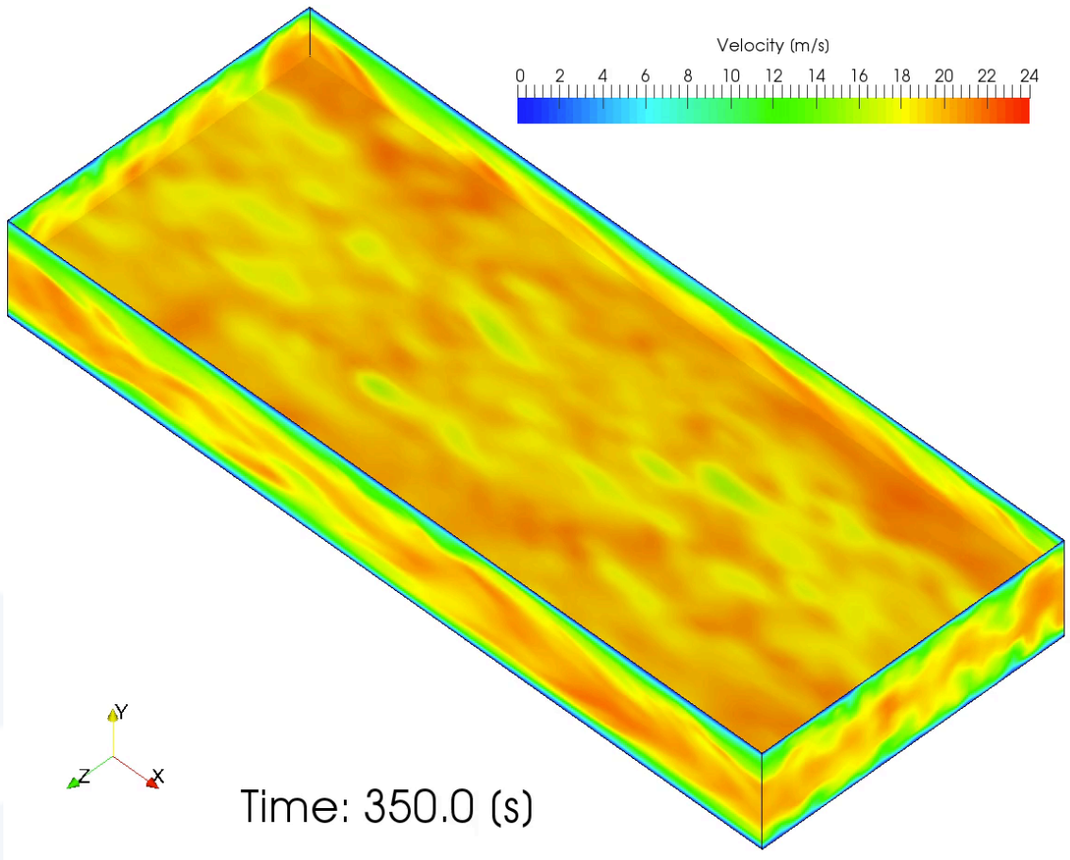


2016年度料金(税込)．EC2とAzureは2015年11月時(NFSサーバ用のインスタンス1台の料金も考慮)

Copyright © 2016 OCAEL All rights reserved.

ベンチマークテスト流れ場(格子数3M)

チャネル流れ ($Re_\tau = 110$)



解析条件

$L_x \times L_y \times L_z = 5\pi \times 2 \times 2\pi$

$Re_\tau = u_\tau \, \delta / \mu \; = 110$ [-]

ここで

u_τ：壁面摩擦速度 [m/s]

δ ：チャネル半幅 [m] ($=L_y / 2$)

μ ：動粘性係数 [m^2/s^2]

主流方向(x): 一定の圧力勾配

主流方向(x), スパン方向(z): 周期境界

乱流モデル: 無し(laminar)

速度線型ソルバ: BiCG (前処理DILU)

圧力線型ソルバ: CG (前処理DIC)

(RapidCFDでは前処理はAINV[1])

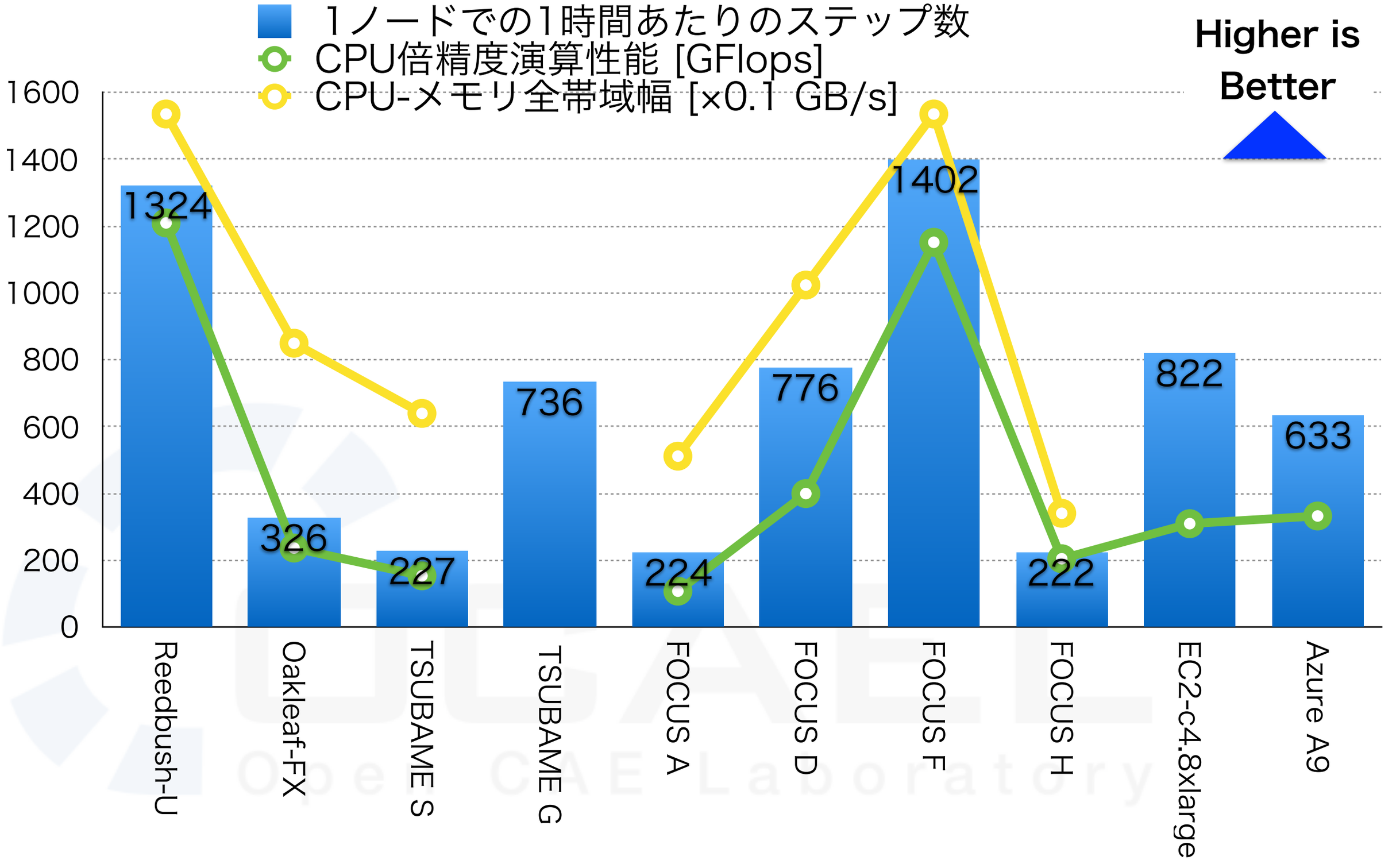
領域分割手法：scotch

OpenFOAMのバージョンは2.3.0.　TSUBME G(GPU)ではGPU版のRapidCFDを用いた.

[1] Algorithm for Sparse Approximate Inverse Preconditioners in the Conjugate Gradient Method, Ilya B. Labutin, Irina V. Surodina

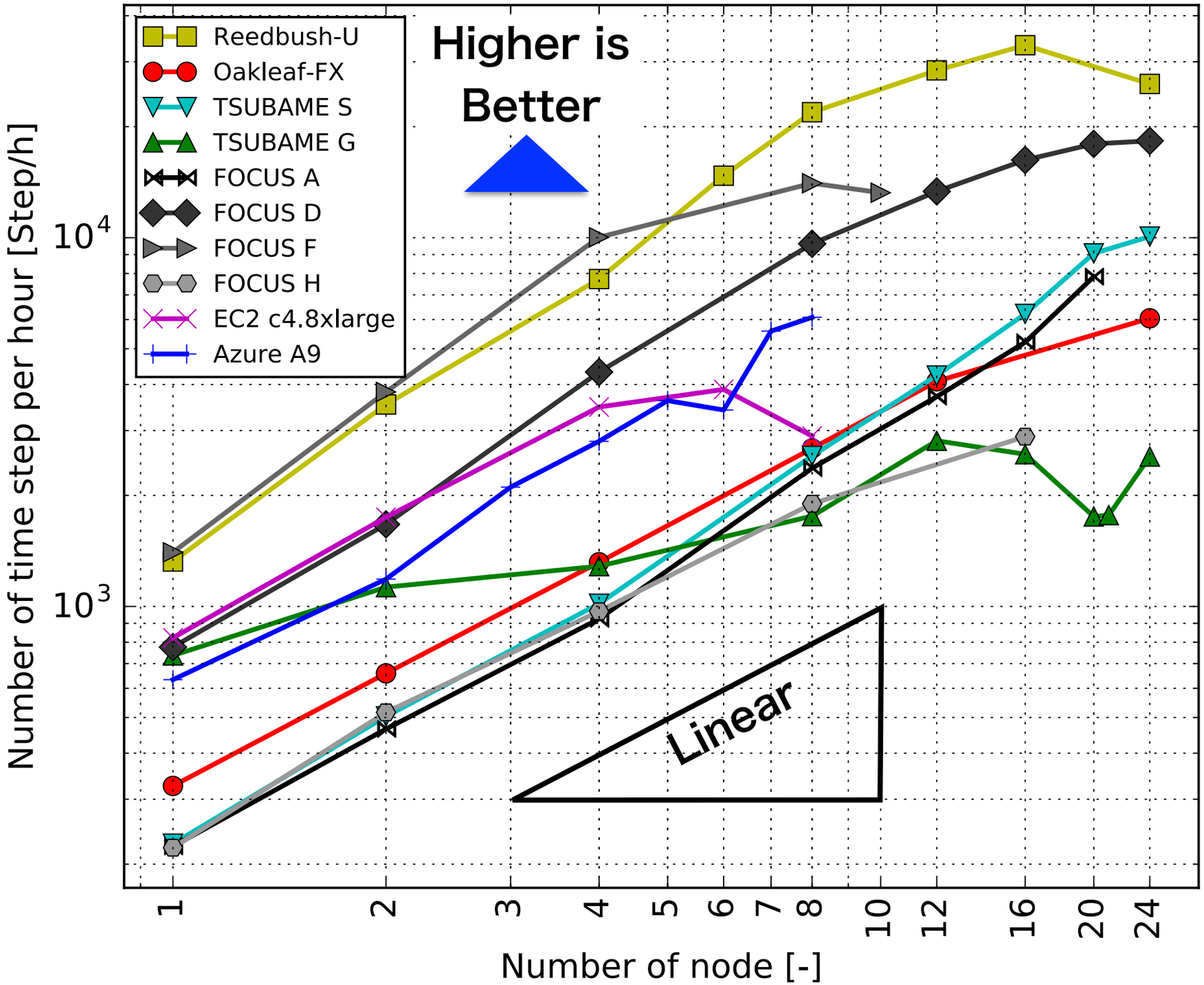
Copyright © 2016 OCAEL All rights reserved.

ステップ数・CPU演算性能・メモリ帯域幅



Copyright © 2016 OCAEL All rights reserved.

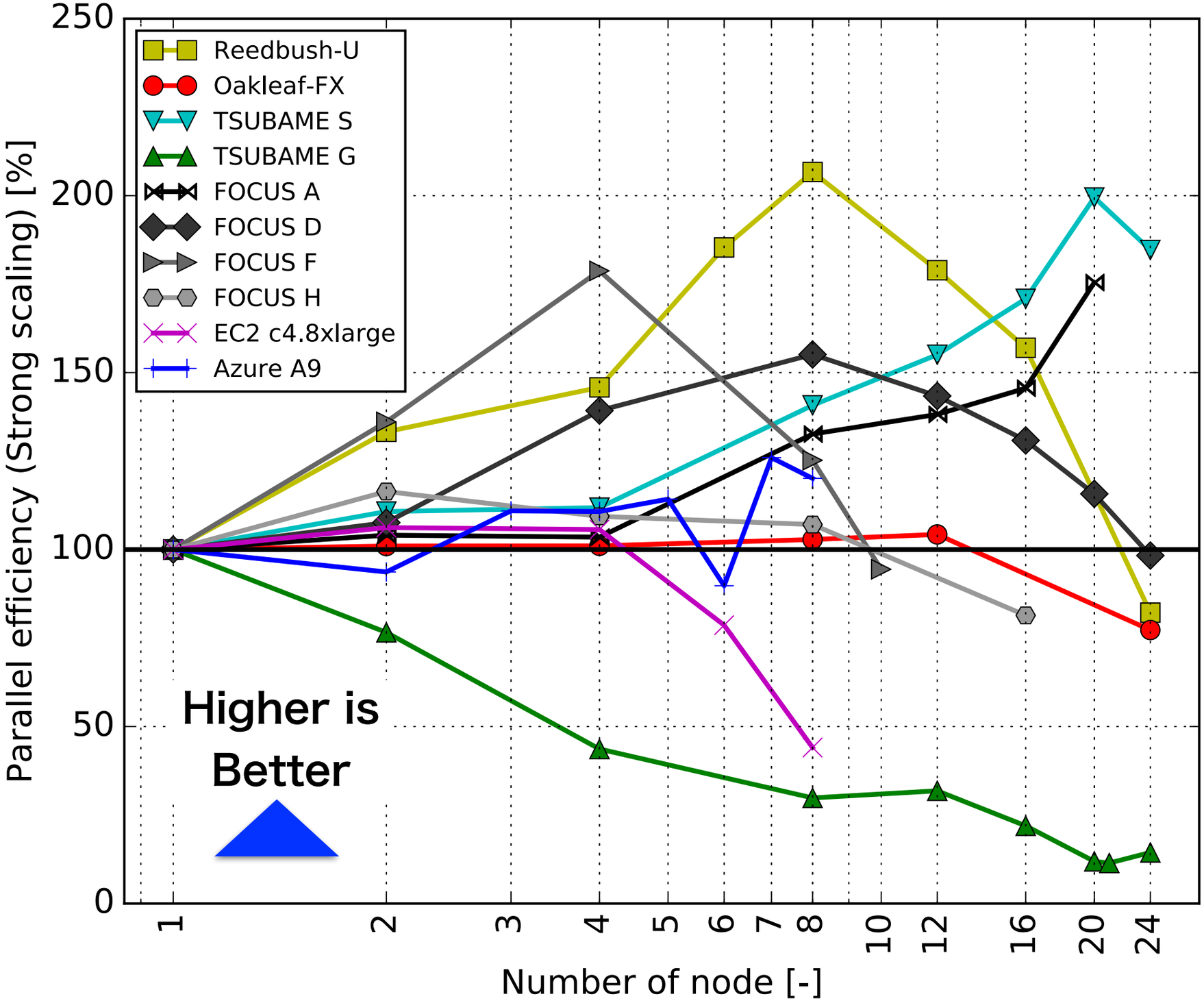
1時間あたりのステップ数



- 最高速
  - ✓ 4ノードまで: FOCUS F
  - ✓ 6ノード以上: Reedbush-U
- 飽和ノード数
  - ✓ FOCUS F: 8
  - ✓ Reedbush-U: 16
  - ✓ EC2: 6
  - ✓ TSUBAME-G: 12

Copyright © 2016 OCAEL All rights reserved.

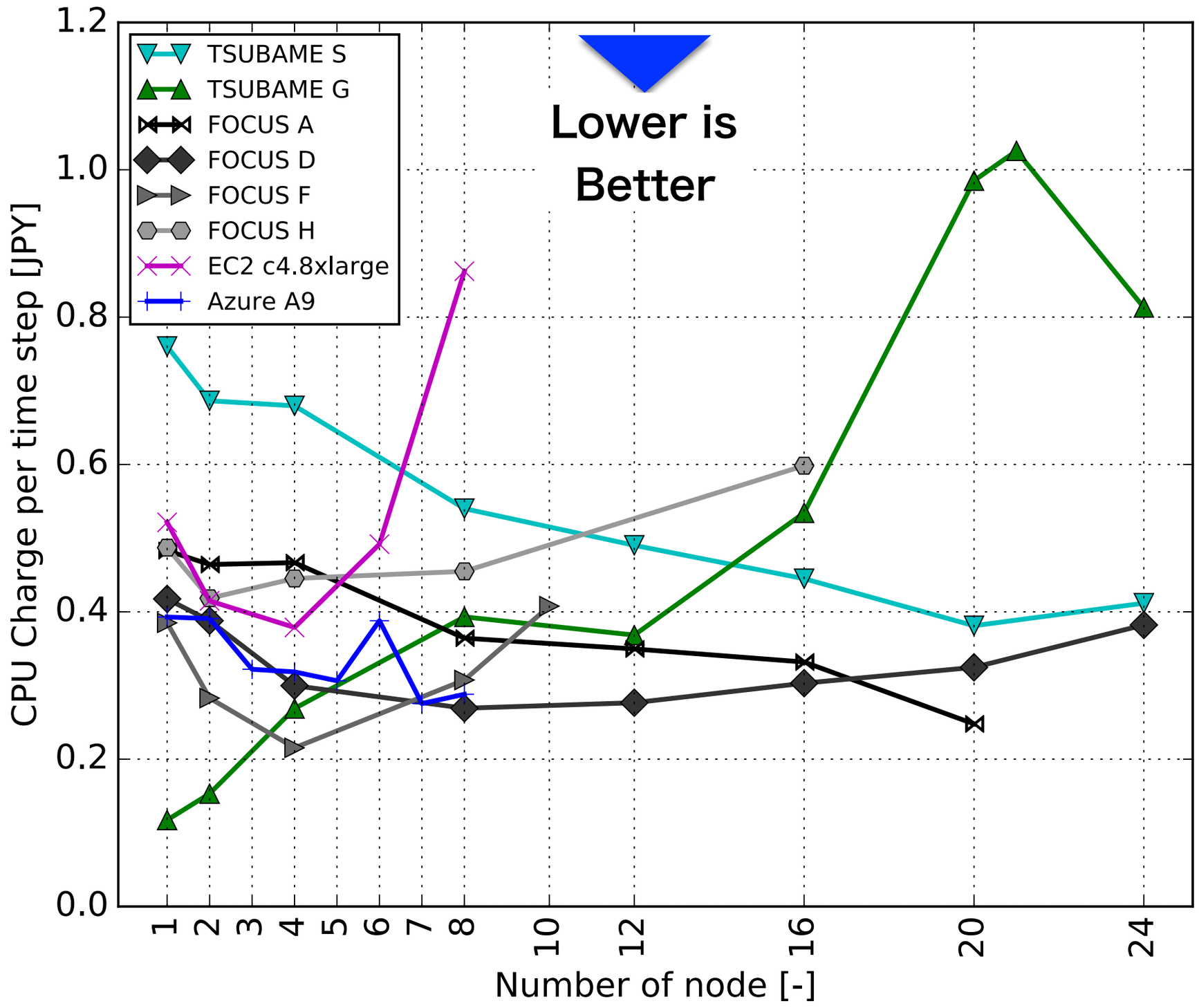
並列化効率(Strong scaling)



- SPARC64 CPU機のOakleaf-FXは12ノードまでほぼリニア
- Intel CPU機は概ねスーパリニア
- ピーク効率のノード数
  - ✓ FOCUS F: 4
  - ✓ Reedbush-U: 8
  - ✓ FOCUS D: 8
  - ✓ TSUBAME S: 20
- インターコネクトが10GbEであるEC2は6ノード以上で劣化
- TSUBAME Gは多ノードで急激に性能劣化

Copyright © 2016 OCAEL All rights reserved.

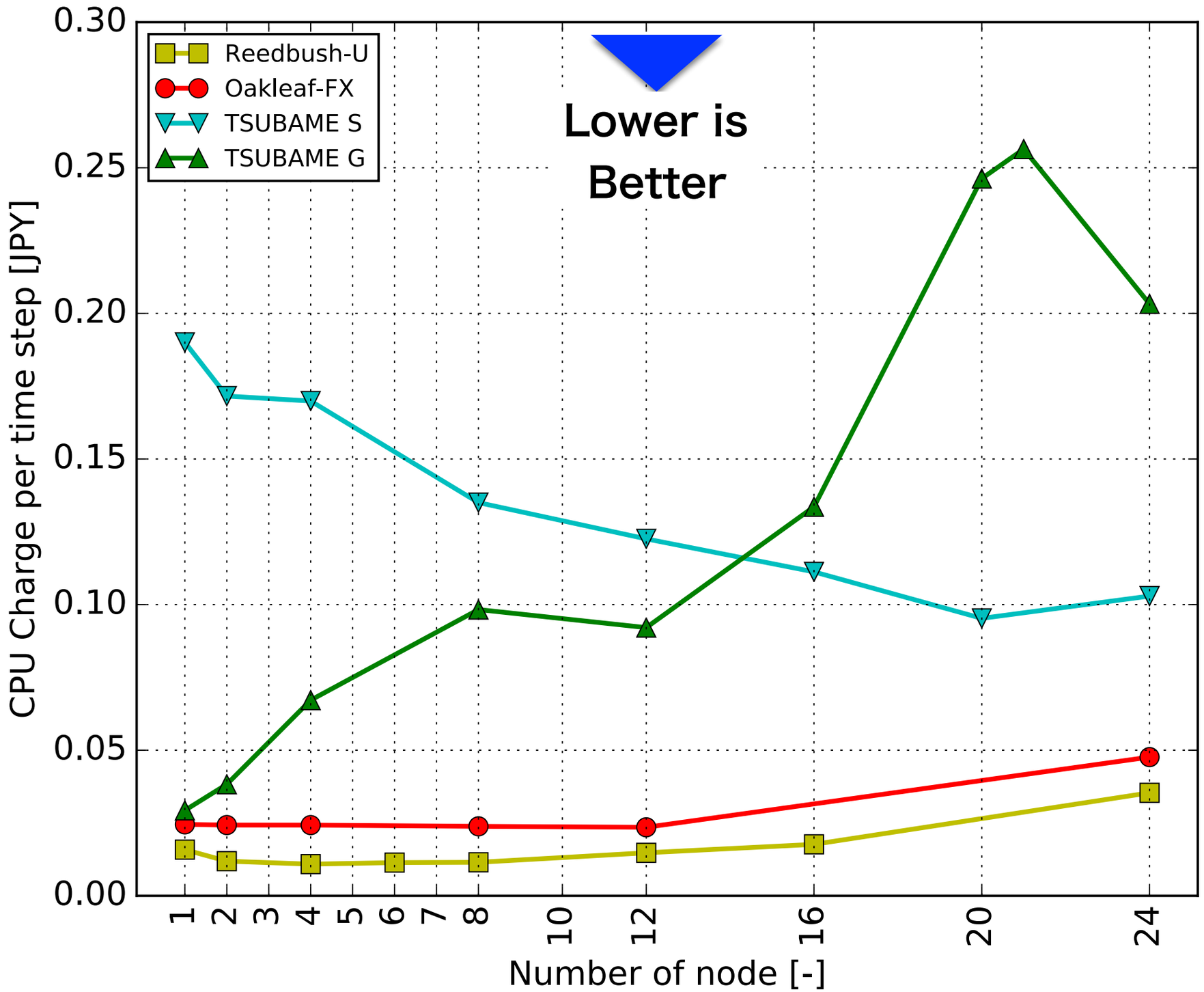
ステップ毎の課金(成果非公開型)



- 最安価
  - ✓ 1, 2ノード: TSUBAME G
  - ✓ 4ノード: FOCUS F
  - ✓ 8~16, 24ノード : FOCUS D
  - ✓ 20ノード: FOCUS A
- 基本的に課金額は並列化効率に反比例するが, FOCUSでは多ノード割引率にも依存する.

Copyright © 2016 OCAEL All rights reserved.

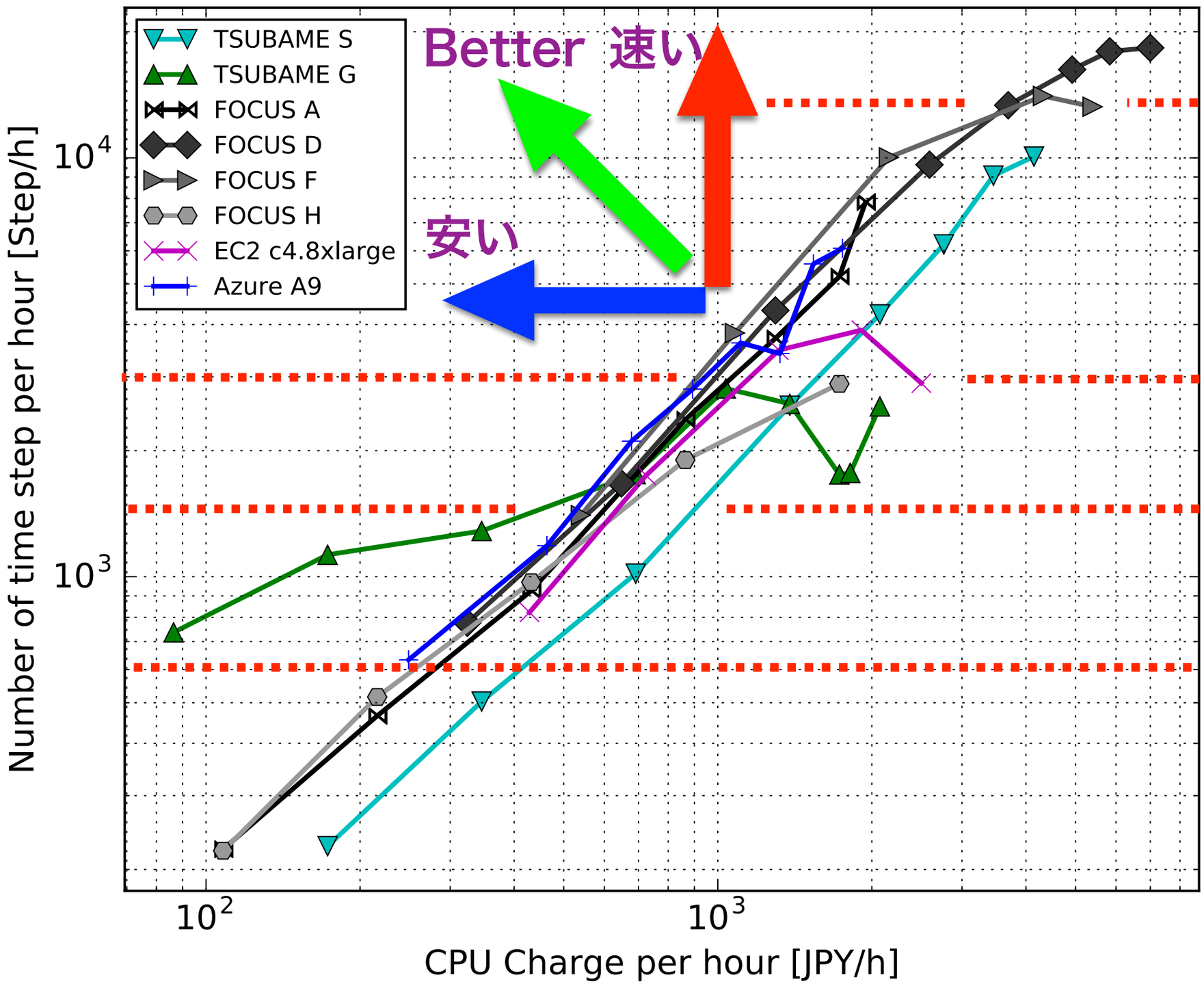
ステップ毎の課金(成果公開型)



- 最安価
  - ✓ 全ノード: Reedbush-U
  - ✓ Oakleaf-FXが次点
- Reedbushは4ノード，Oakleaf-FXは12ノードを越えたノード利用分には消費係数が2倍となるので，ノード数が増えると次第に高価となる．ただし，申込ノード数に依存する．

Copyright © 2016 OCAEL All rights reserved.

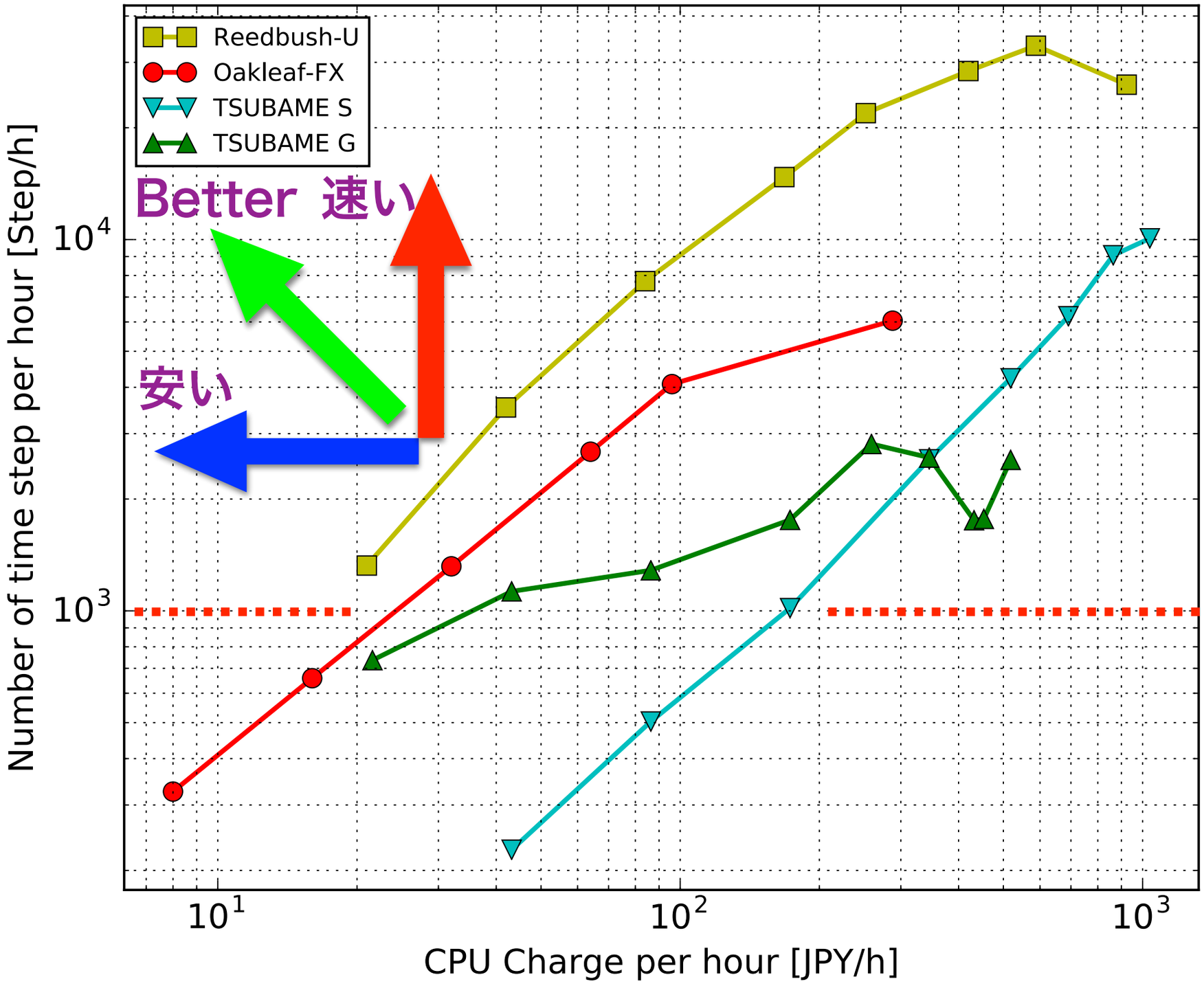
1時間の課金とステップ数(成果非公開型)



ステップ数	最安価
20,000～	FOCUS D
4,000～20,000	FOCUS F
2000～4000	Azure A9
600～2000	TSUBAME G(GPU), Azure A9
～600	FOCUS H

Copyright © 2016 OCAEL All rights reserved.

1時間の課金とステップ数(成果公開型)



ステップ数	最安価
10,00～	Reedbush-U
～1,000	Oakleaf-FX

Copyright © 2016 OCAEL All rights reserved.

まとめ

- 東京大学・東京工業大学・FOCUSのスパコン，Amazon EC2, Microsoft Azureのクラウドにおいて，格子数3MのチャンネルフローによるOpenFOAMのベンチマークテストを実行した.
- 1時間で実行可能なステップ数と課金額の関係から，ステップ数の領域毎に安価なシステムが分かれる結果となった.
- 計測データおよび計測用ファイル一式は，オープンCAE学会V&V委員会のGithubレポジトリで公開しているので，興味があればご参照ください.

https://github.com/opencae/OpenFOAM-BenchmarkTest

謝辞

東京工業大学学術国際情報センター共同利用推進室の佐々木様から，OpenFOAMとRapidCFDの評価用として，TSUBAME 2.5の計算機リソースを提供して頂きました. 日本Microsoft(計測当時)の佐々木様から，Microsoft Azure A9でのベンチマークの結果を提供して頂きました. 電通国際情報サービス(計測当時)の住友様には，Amazon EC2のベンチマーク結果を提供して頂きました. 青子守歌様には，RapidCFDのビルド及びベンチマークについてご協力頂きました. ここに深く感謝致します.

Copyright © 2016 OCAEL All rights reserved.

(附録)ノード時間料金の算出条件

システム	成果公開	ノード時間料金[円](※1)	ノード時間料金の算出条件
Reedbush-U	公開	21.5	最も高価となるグループコース(4ノード，企業)，利用期間1ヶ月間
Oakleaf-FX		9	最も高価となるグループコース(12ノード，企業)，利用期間1ヶ月間
TSUBAME S		43.2	成果非公開は成果公開の4倍の料金．GはSの1/2の料金．従量利用，最大計算時間：1時間，優先度：標準，実行時間ごとの係数：1の場合．学内・共同研究利用(ノード時間料金10円，40円)は除外
TSUBAME G		21.6	
TSUBAME S	非公開	172.8	
TSUBAME G		86.4	
FOCUS A		108	多ノード割引きを考慮
FOCUS D		324	
FOCUS F		540	
FOCUS H		108	
EC2-c4.8xlarge		295.8	計測時：2015年11月15日．リージョン：東京．NFSサーバ: 132.8円/h(c3.4xlarge)(※2)
Azure A9		214.8	計測時：2015年11月25～26日．リージョン：West US．NFSサーバ: 33.3円/h(D3)(※2)
(※1)税込．計測時明記以外は2016年度料金 (※2)NFSサーバ用のインスタンス1台の料金も考慮			

Copyright © 2016 OCAEL All rights reserved.

(附録)ソフトウェア・コンパイラ・MPIライブラリ

システム	ソフトウェア	コンパイラ(※1)	MPI
Reedbush-U	OpenFOAM 2.3.0	Gcc-4.8.5	OpenMPI 1.8.3
Oakleaf-FX		FCC GM-1.2.1-09	FJMPI GM-1.2.1-09
TSUBAME S		Gcc-4.8.4	OpenMPI 1.6.5(※3)
TSUBAME G(GPU)	RapidCFD (※2)	nvcc (cuda-6.5)	OpenMPI 1.8.4(※4)
FOCUS A, D, F, H	OpenFOAM 2.3.0	Gcc 4.8.3	OpenMPI 1.6.5(※3)
EC2 c4.8xlarge		Gcc 4.8.5	OpenMPI 1.8.5(※5)
Azure A9		Gcc 4.8.3	Intel MPI 5.1.1.109(※6)

※1) 最適化フラグ: -O3 ※2) rev: d3733257dee5fb9999b918f5c26a1493cebb603c ※3) mpirunオプション: -bind-to-core -mca btl openib,sm,self ※4) mpirunオプション: -bind-to core -mca btl openib,sm,self ※5) mpirunオプション: -bind-to core ※6) Azure A9のLinux OSでは，MPI通信にRDMAを使うためにはIntel MPIが必要

Copyright © 2016 OCAEL All rights reserved.

(附録)検討ノード数・MPI数

システム	コア /ノード	設定MPI数 /ノード	解析ノード数	解析MPI数 (フラットMPI)
Reedbush-U	36	←	1, 2, 4, 6, 8,12, 16, 24	36～864
Oakleaf-FX	16	←	1, 2, 4, 8, 12, 24(※1)	16～384
TSUBAME S	12	10(※2)	1, 2, 4, 8, 12, 16, 20, 24 (FOCUS Aは20まで)	10～240
TSUBAME G	3GPU	←		3～72
FOCUS A	12	10(※2)		10～200
FOCUS D	20	←		20～480
FOCUS F	40	←	1, 2, 4, 8, 10	40～400
FOCUS H	8	←	1, 2, 4, 8, 16	8～128
EC2 c4.8xlarge	18	←	1, 2, 4, 6, 8	18～144
Azure A9	16	←	1, 2, 3, 4, 5, 6, 7, 8	18～128

※1) 12ノード以上は，TOFU単位である12ノードの倍数で検討した

※2) 事前の検討により12コアを使用するより10コア使用のほうが計算が速かった

Copyright © 2016 OCAEL All rights reserved.

(附録)ステップ数・演算性能・メモリ帯域幅

システム	倍精度 理論演算性能 [GFlops] (F)	ノードあたりの CPU(GPU)-メモリ 全帯域幅[GB/s] (B)	1ノードでの 1時間あたりの ステップ数 (S)	B/F 値	S/F 値	S/B 値
Reedbush-U	1210	153.6	1324	0.13	1.09	8.62
Oakleaf-FX	237	85	326	0.36	1.38	3.84
TSUBAME S	154	64	227	0.42	1.47	3.55
TSUBAME G	3930	450	736	0.11	0.19	1.64
FOCUS A	108	51.2	224	0.47	2.07	4.38
FOCUS D	400	102.4	776	0.26	1.94	7.58
FOCUS F	1152	153.6	1402	0.13	1.22	9.13
FOCUS H	205	34.1	222	0.17	1.08	6.51
Amazon EC2	310	不明	822	-	2.65	-
Azure A9	333	不明	633	-	1.90	-

Copyright © 2016 OCAEL All rights reserved.