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（OCAEL，北海道大学招へい教員，東京大学客員研究員）
住友正紀（電通国際情報サービス）

Copyright © 2015 OCAEL All rights reserved.

研究の背景

- **大学のスパコン：東京大学FX10，東京工業大学TSUBAME, Etc.**
  - ✓ 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年単位での一口タイプ課金

- **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, Etc.**
  - ✓ 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 © 2015 OCAEL All rights reserved.

対象システムの特徴

- **東京大学 Oakleaf-FX**
  - ✓ 利用には課題申請が必要
  - ✓ 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**
  - ✓ 法人のみ使用可能．課題申請は不要．使った分だけ課金
  - ✓ 2015年度下半期は期間占有が多く，従量利用が混雑
- **Amazon EC2 (クラウド)**
  - ✓ 使った分だけ課金(だだし，課金は1時間単位)
  - ✓ 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)**
  - ✓ 使った分だけ課金(分単位)
  - ✓ 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 © 2015 OCAEL All rights reserved.

対象システムの性能

システム	CPU [GPU]	周波数 [GHz]	コア/ノード	インターコネクト
東京大学 Oakleaf-FX	Fujitsu SPARC64 IXfx	1.848	16コア	Tofu, 40Gbps ×双方向 ×10ポート
東京工業大学 TSUBAME 2.5 Sキュー(※1) Gキュー(GPUのみ)	Intel Xeon E5-2670	2.93	6コア × 2 CPU	Infiniband-QDR, 40Gbps×2
	[nVIDIA Tesla K20X]	—	[3GPU]	
FOCUS Dシステム	Intel Xeon E5-2670 v2	2.5	10コア × 2 CPU	Infiniband-FDR, 56Gbps
Amazon EC2 c4.8xlarge(※2)	Intel Xeon E5-2666 v3	2.9	9コア × 2 CPU	10GbE, 10Gpps
Microsoft Azure A9(※2)	Intel Xeon E5-2670	2.6	8コア × 2CPU	InfiniBand-QDR, 40Gbps
(※1)ターボブースト有効 (※2)仮想マシン．Hyper-threading無効				

現時点で各機関での最高性能のシステムを対象

Copyright © 2015 OCAEL All rights reserved.

ノード時間料金

システム	ノード時間料金	備考
Oakleaf-FX	20.9円(成果公開)	最も高いノード時間料金となるグループコース(企業)，利用期間1ヶ月間の場合．24ノードの申込を想定．大学・公共機関等用のコース(ノード時間料金4.63円，4.82円)は除外.
TSUBAME S	43.2円(成果公開) 172.8円(成果非公開)	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．東京工業大学の学内・共同研究利用(ノード時間料金10円，40円)は除外.
TSUBAME G (GPU)	21.6円(成果公開) 86.4円(成果非公開)	
FOCUS D	324円(成果非公開)	33ノード以上で割引となる．今回は24ノード以下で該当せず.
EC2 c4.8xlarge	295.8円(成果非公開)(※1)	計測時(2015年11月15日)の料金．リージョン：東京．NFSサーバ: 132.8円/h(c3.4xlargeインスタンス)
Azure A9	214.8円(成果非公開)(※1)	計測時(2015年11月25～26日)の料金．Virtual Machine．リージョン：West US．NFSサーバ: 33.3円/h(D3インスタンス)
料金は税込．(※1) 通常(オンデマンド)料金．NFSサーバ用のインスタンス1台の料金も考慮		

産業利用が可能なコース・サービスのみを対象とした

Copyright © 2015 OCAEL All rights reserved.

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
Oakleaf-FX	OpenFOAM 2.3.0	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 D	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)

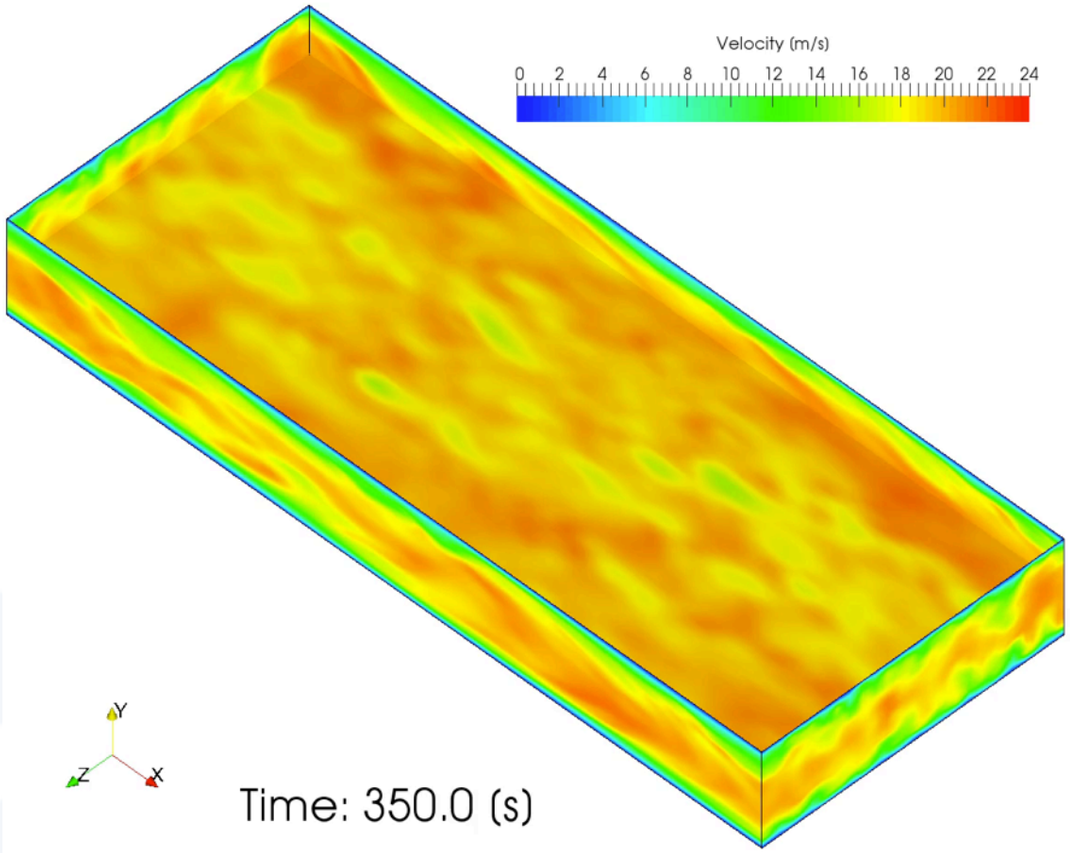
※1) 最適化フラグ: -O3 ※2) rev: d3733257dee5fb9999b918f5c26a1493cebb603c ※3) mpirunオプション: -bind-to-core -mca btl openib,sm,self ※4) mpirunオプション: -bind-to core -mca btl openib,sm,self ※5) mpirunオプション: -bind-to core ※6) Azure A9のLinux OSでは，MPI通信にRDMAを使うためにはIntel MPIが必要

OpenFOAMやRapidCFDのソースや最適化フラグはデフォルトのまま

Copyright © 2015 OCAEL All rights reserved.

ベンチマークテストの流れ場

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)

時間刻み: 0.002 [s]

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])

格子数: 約3M (240(x)×130(y)×96(z))

領域分割手法：scotch

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

Copyright © 2015 OCAEL All rights reserved.

時間ステップ毎の平均実行時間

ソルバのログ

(各種初期化，設定・格子・初期値のファイル入力)

```
Time = 0.002 最初(1回目)の時間ステップ
ExecutionTime = 1.35 s  ClockTime = 3 s
:
:
Time = 0.102 最終の前(51回目)の時間ステップ
ExecutionTime = 11.18 s  ClockTime = 13 s
```

(計算結果のファイル出力)

```
Time = 0.104 最終(52回目)の時間ステップ
ExecutionTime = 11.37 s  ClockTime = 15 s

End
```

本ベンチマークテストでの時間ステップ毎の平均実行時間の計算方法

```
( 51回目のClockTime
-  1回目のClockTime)/50
```

計算例

```
(13-3)/50=10/50=0.2 [s]
```

- ソルバのログでのClockTimeはファイルIOやMPI通信などを含む，実際のソルバの実行時間
- 最初の各種初期化・ファイル入力と最後のファイルの出力を除外
- 50ステップでの平均により，平均実行時間はある程度収束

5回以上ソルバを実行し，最速5位までの実行時間の平均を取得
(ただし，Azureは1回のみ計測)

Copyright © 2015 OCAEL All rights reserved.

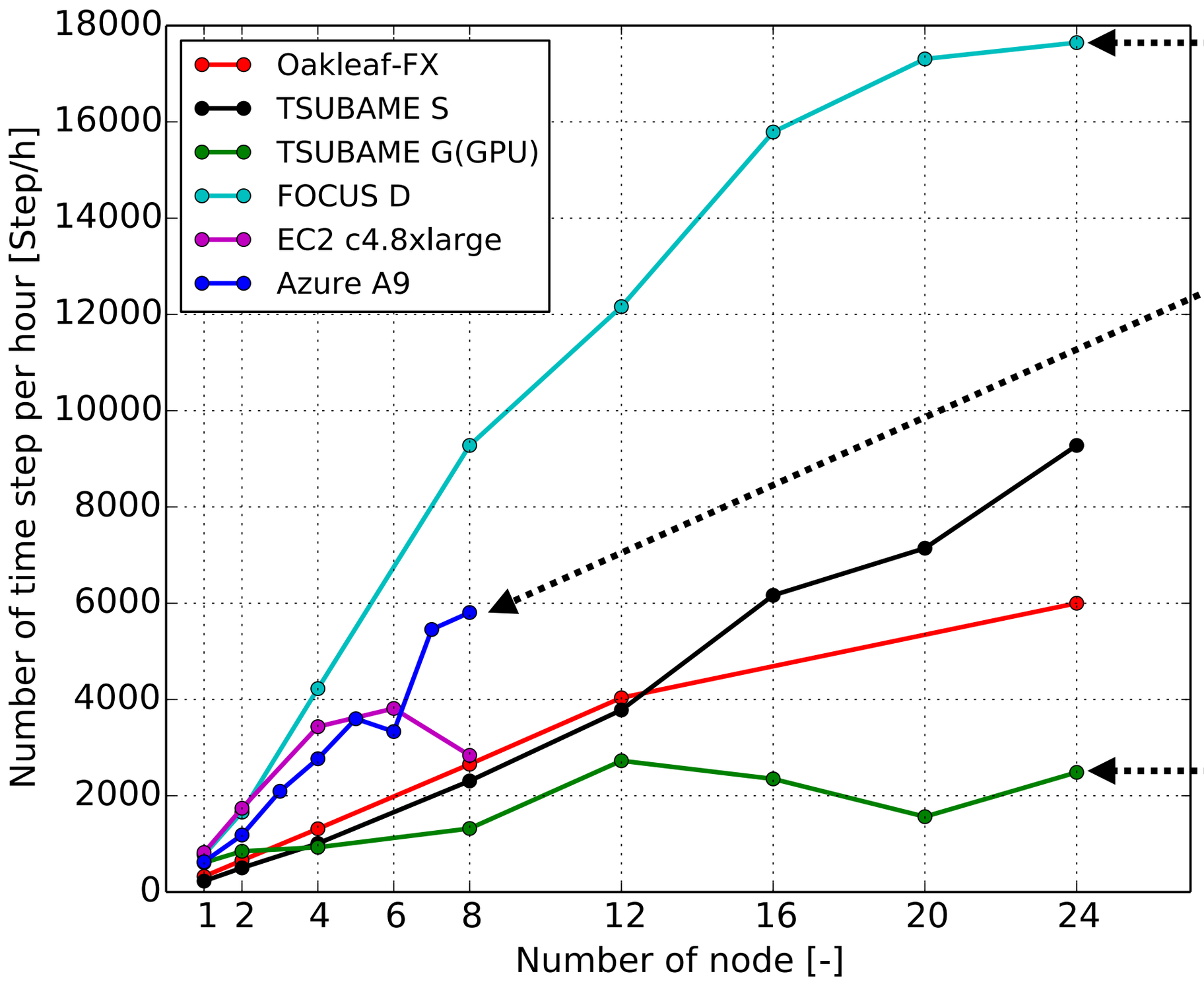
検討ノード数・MPI数

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)
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	10～240
TSUBAME G(GPU)	3 [GPU]	←		3～72
FOCUS D	20	←		20～480
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ため

Copyright © 2015 OCAEL All rights reserved.

1時間あたりの時間ステップ数(全体)



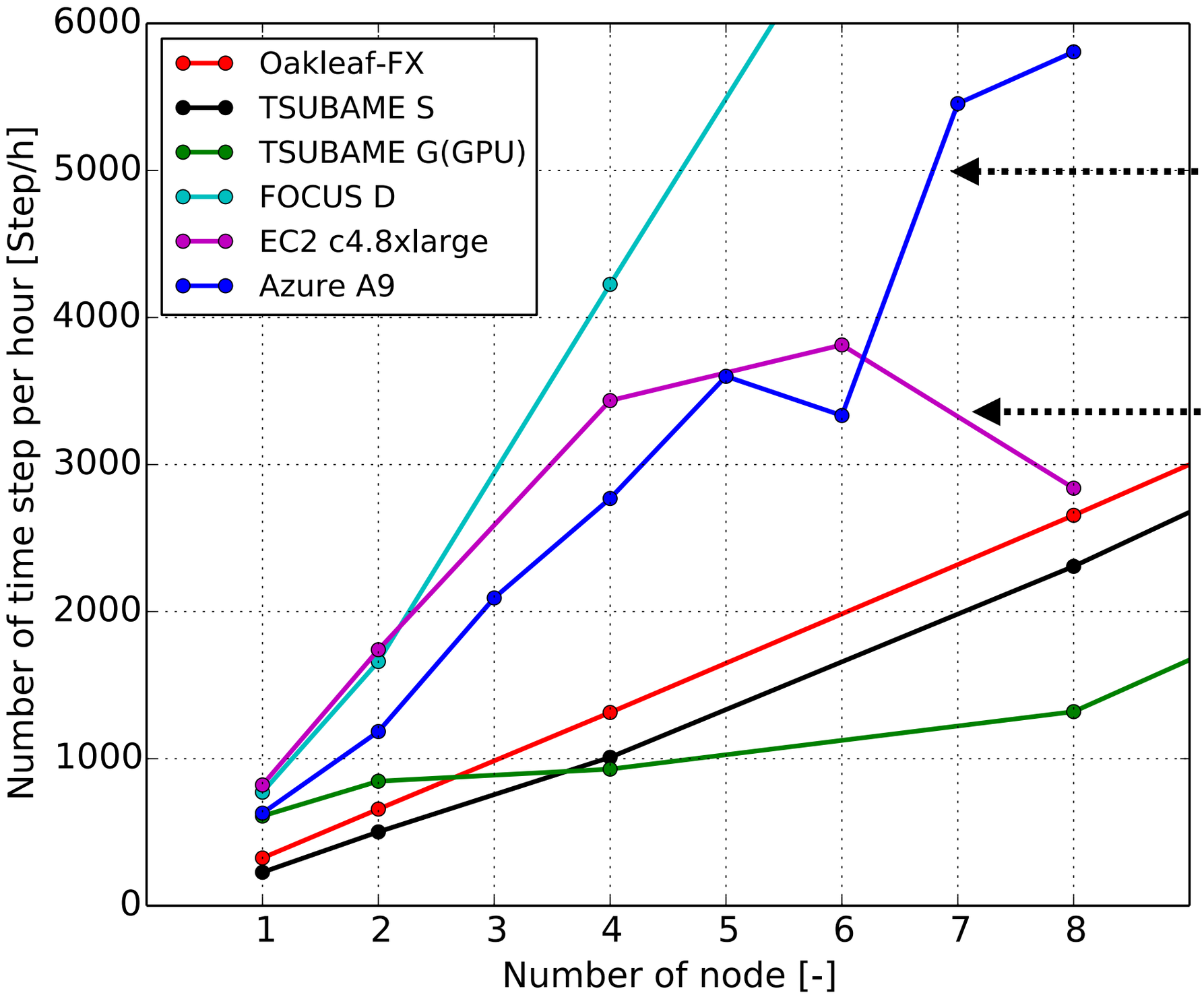
FOCUS Dの性能が最も高い

次に，AzureやEC2のクラウド群の性能が高い

TSUBAME G(GPU)は多ノードで性能が低い

Copyright © 2015 OCAEL All rights reserved.

1時間あたりの時間ステップ数(8ノード以下)

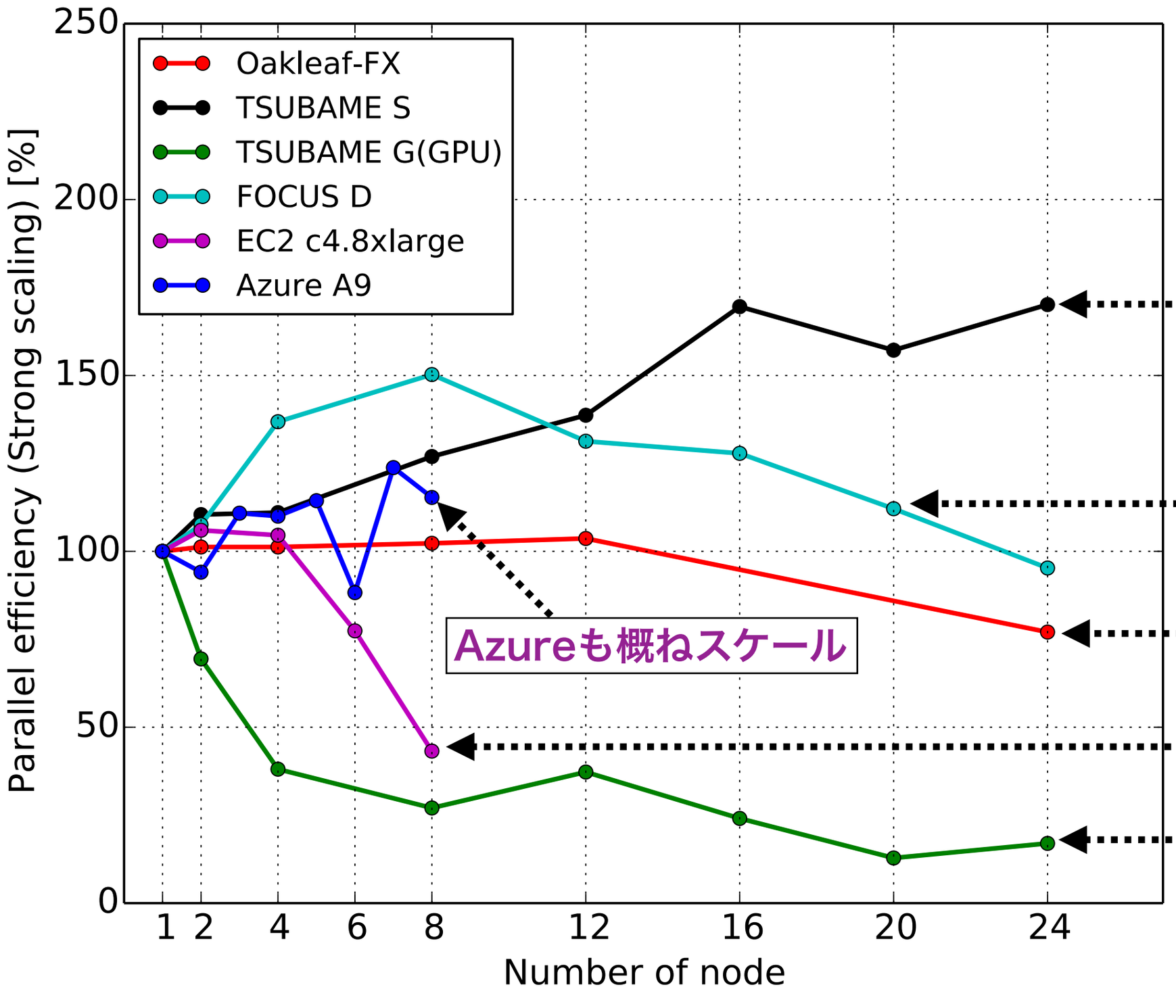


InfiniBandのAzureは6ノードを除き，性能が高い
(1回のみ計測が原因?)

10GbEのEC2は6ノード以上で性能が落ちる

Copyright © 2015 OCAEL All rights reserved.

並列化効率(Strong scaling)



TSUBAME Sは
スーパースケール

FOCUS Dは20ノード
までスケール

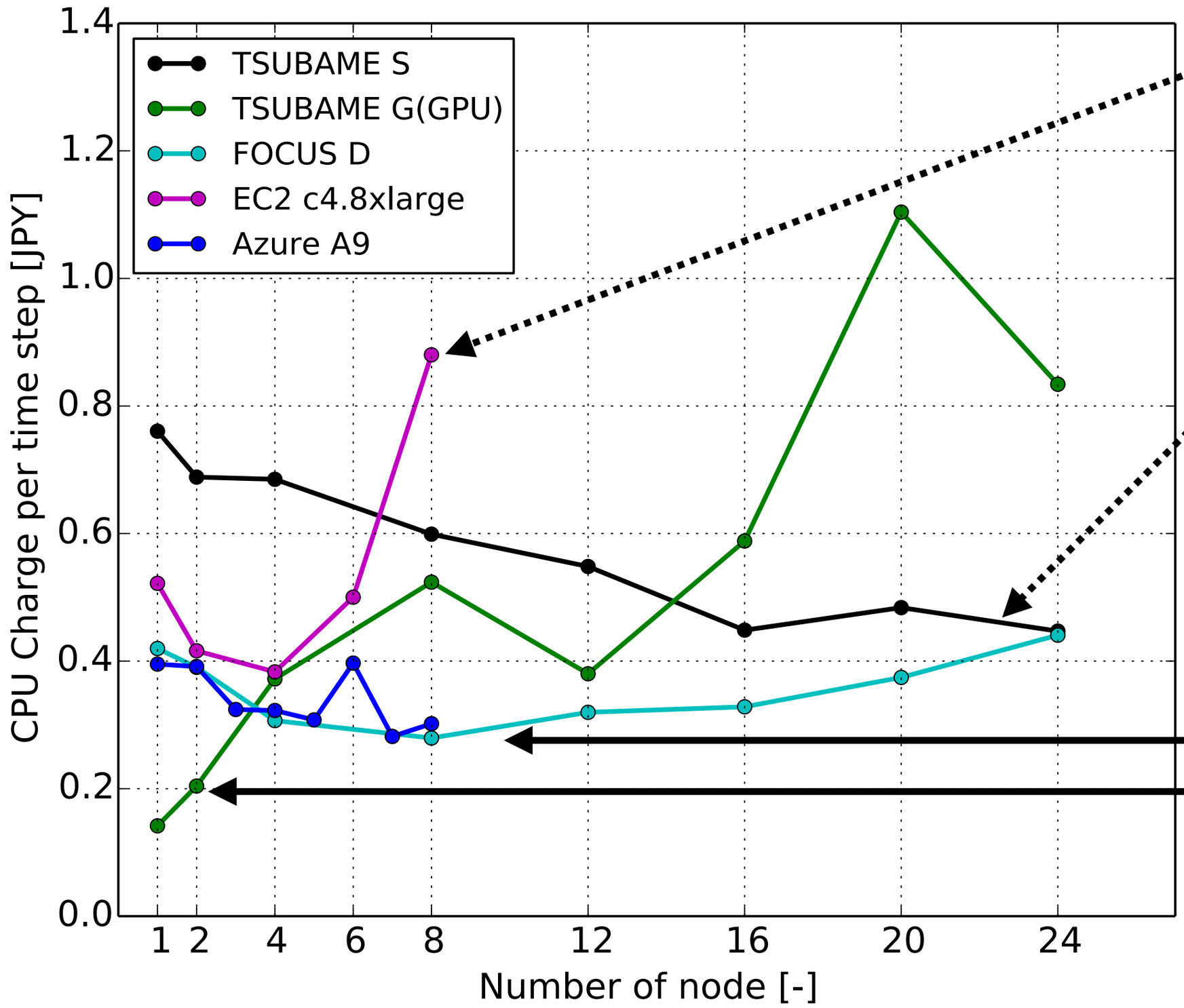
FXは12ノードまで
スケール

10GbEのEC2は6ノー
ド以上で急激に落ちる

TSUBAME G(GPU)は
2ノード以上で並列化
効率が悪い

Copyright © 2015 OCAEL All rights reserved.

時間ステップ毎の課金(成果非公開型)



EC2は6ノード以上で並列化効率が急激に落ちるため高価となる

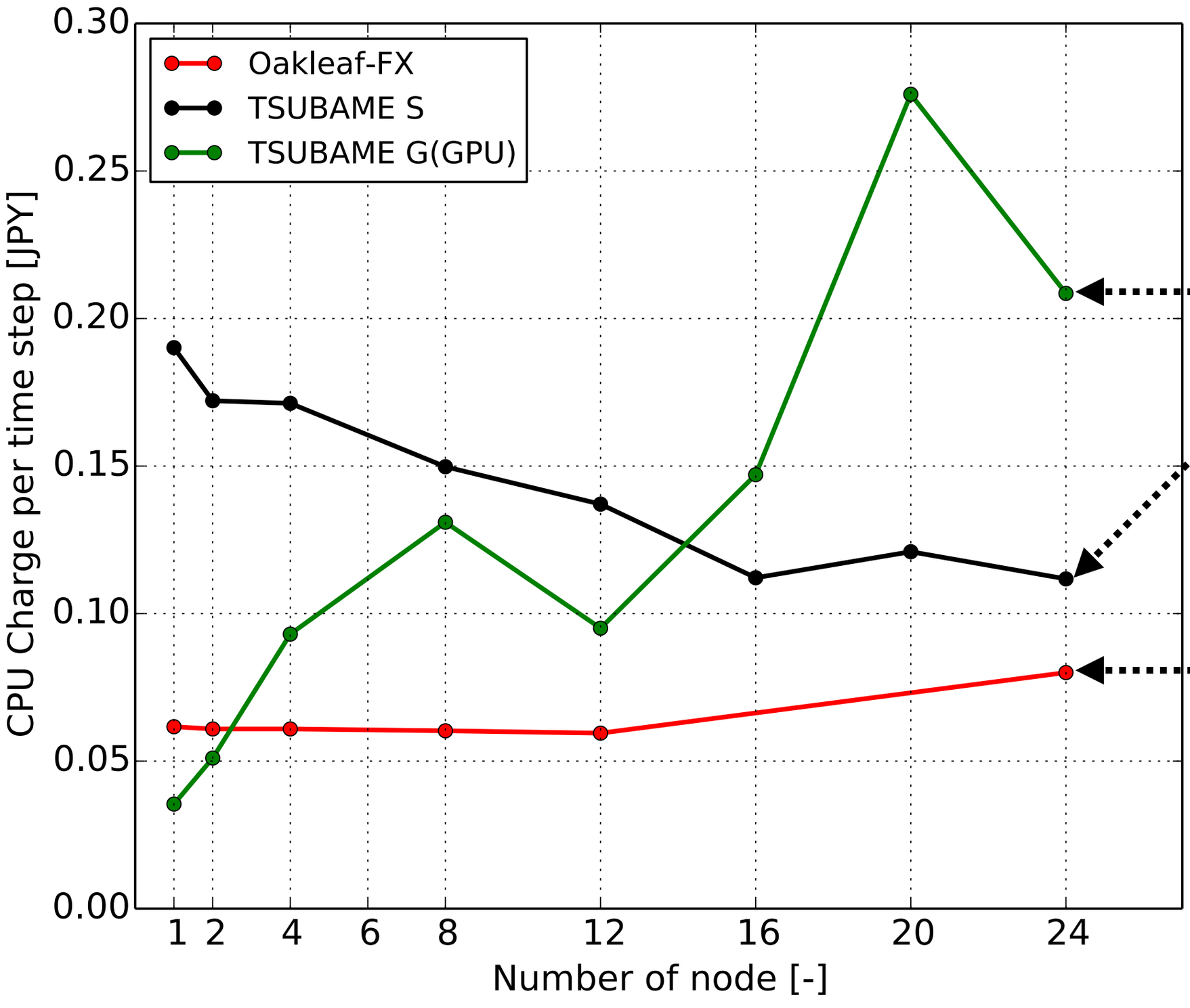
TSUBAME Sは全ノードでスーパースケールしているので，ノード数が増えると安くなる

FOCUS DとAzureは総じて安価

TSUBAME G(GPU)は，2ノードまで安価だが，並列化効率が悪いため多ノードでは高価

Copyright © 2015 OCAEL All rights reserved.

時間ステップ毎の課金(成果公開型)

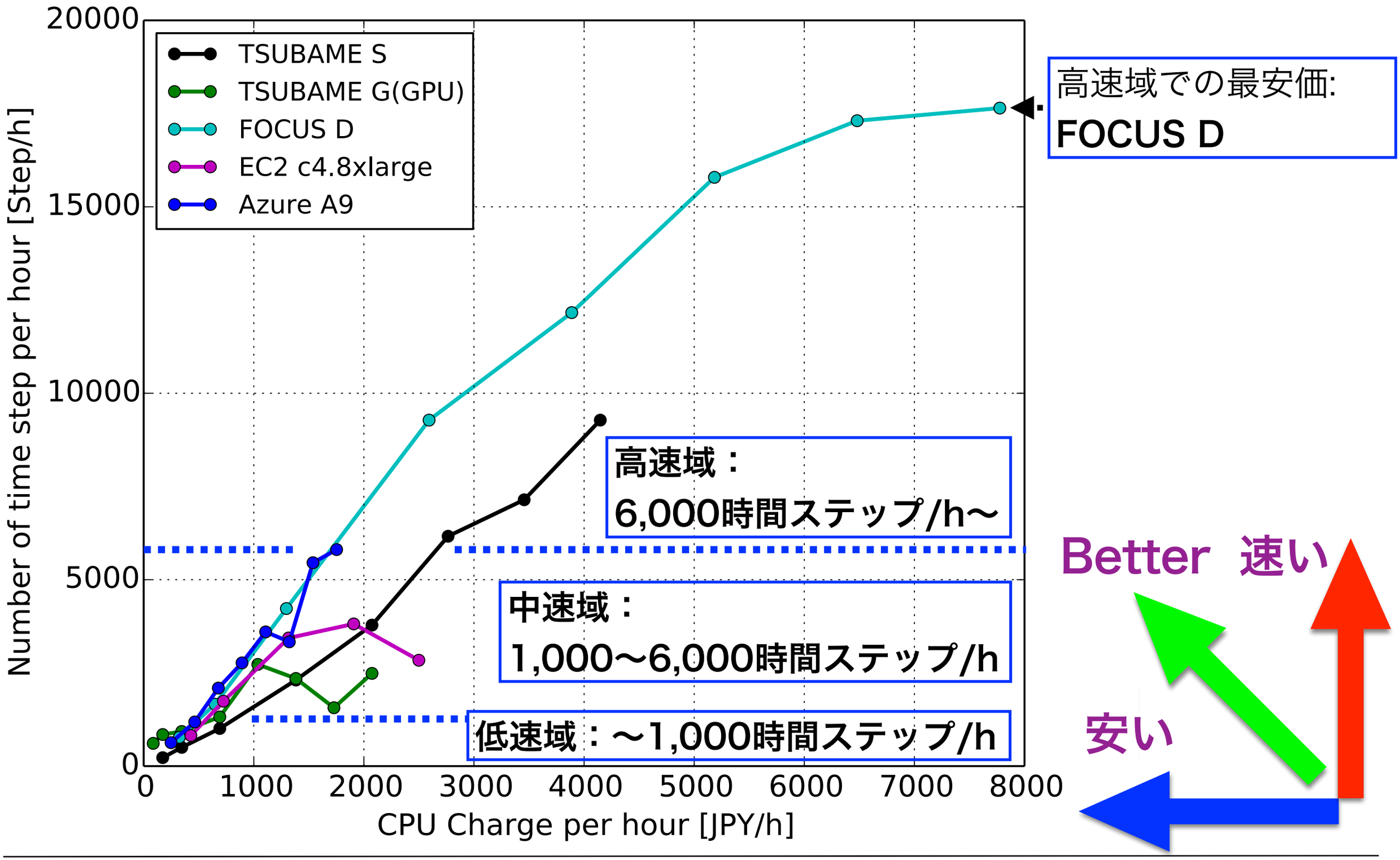


TSUBAME SとG(GPU)は，課金が成果非公開型の1/4になったのみで，成果非公開型と同傾向

FXは12ノードまでスケールしており，ほぼ一定．元々ノード時間料金が安価なので，全体的に安価

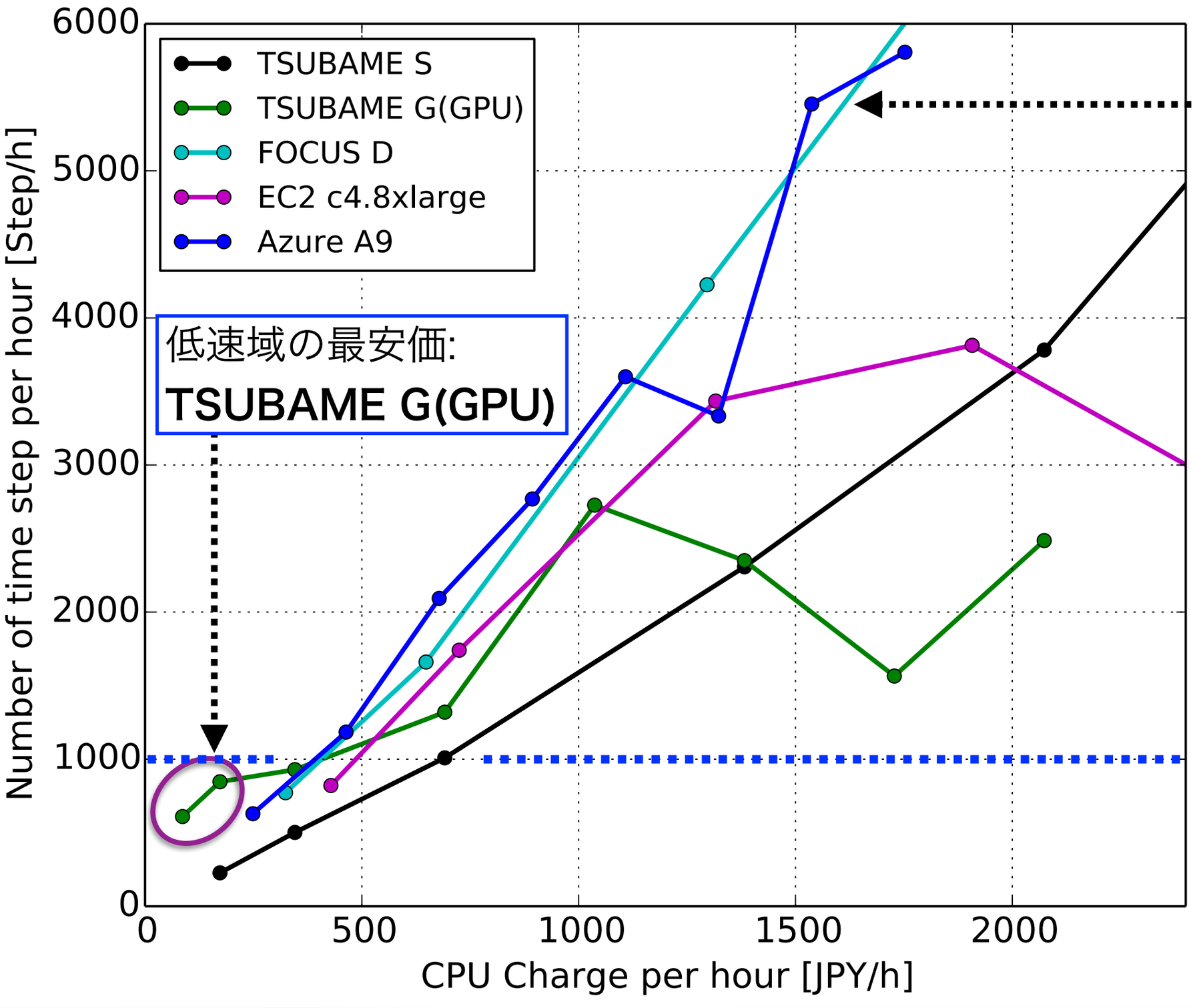
Copyright © 2015 OCAEL All rights reserved.

1時間の課金と時間ステップ数(成果非公開型)



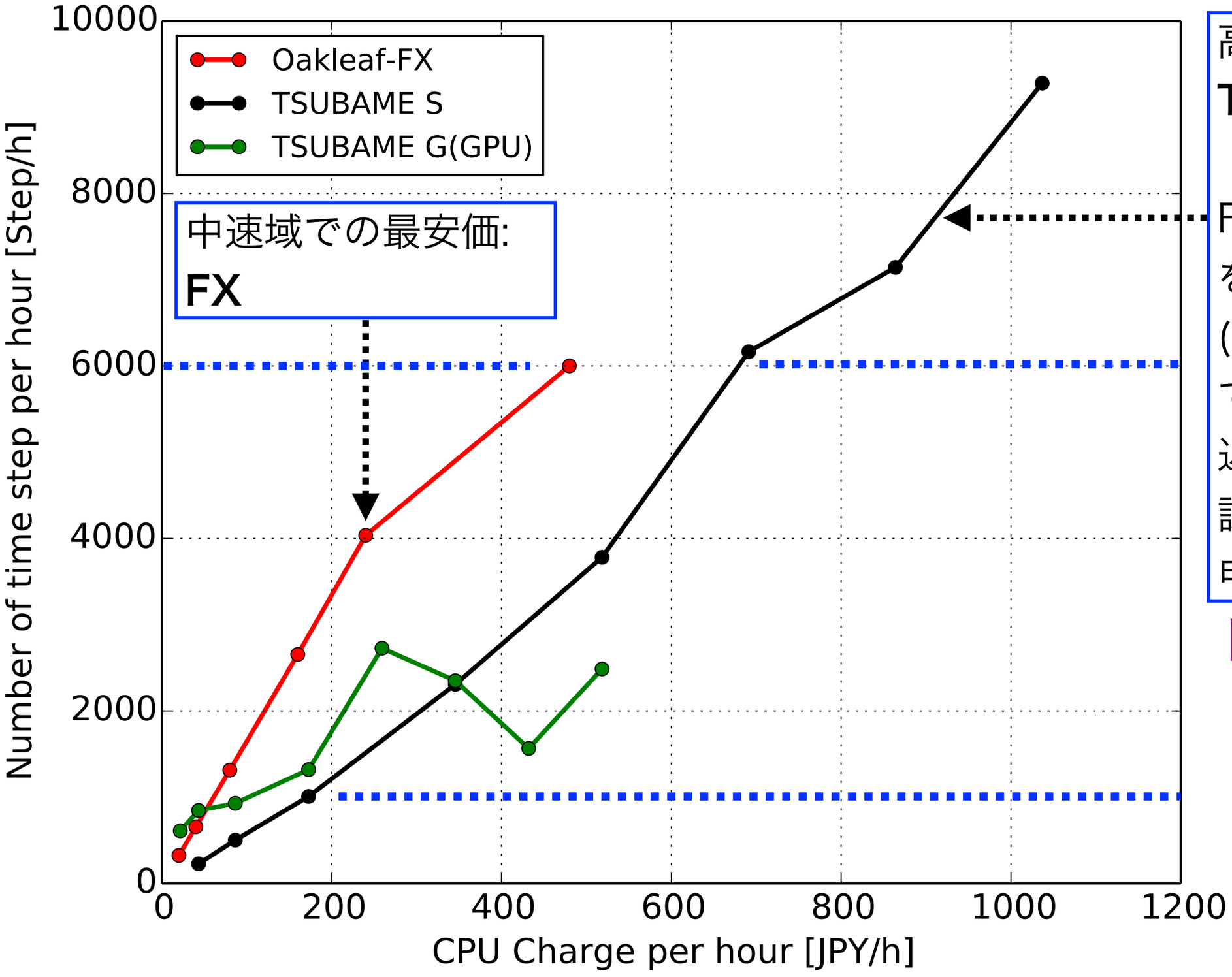
Copyright © 2015 OCAEL All rights reserved.

1時間の課金と時間ステップ数(成果非公開型)



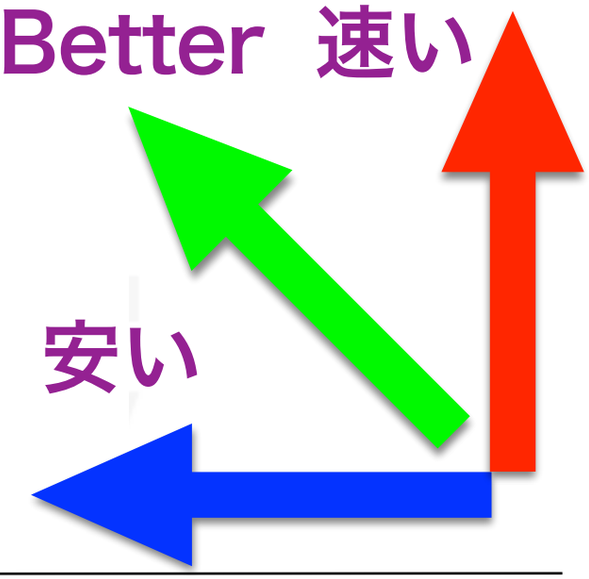
Copyright © 2015 OCAEL All rights reserved.

1時間の課金と時間ステップ数(成果公開型)



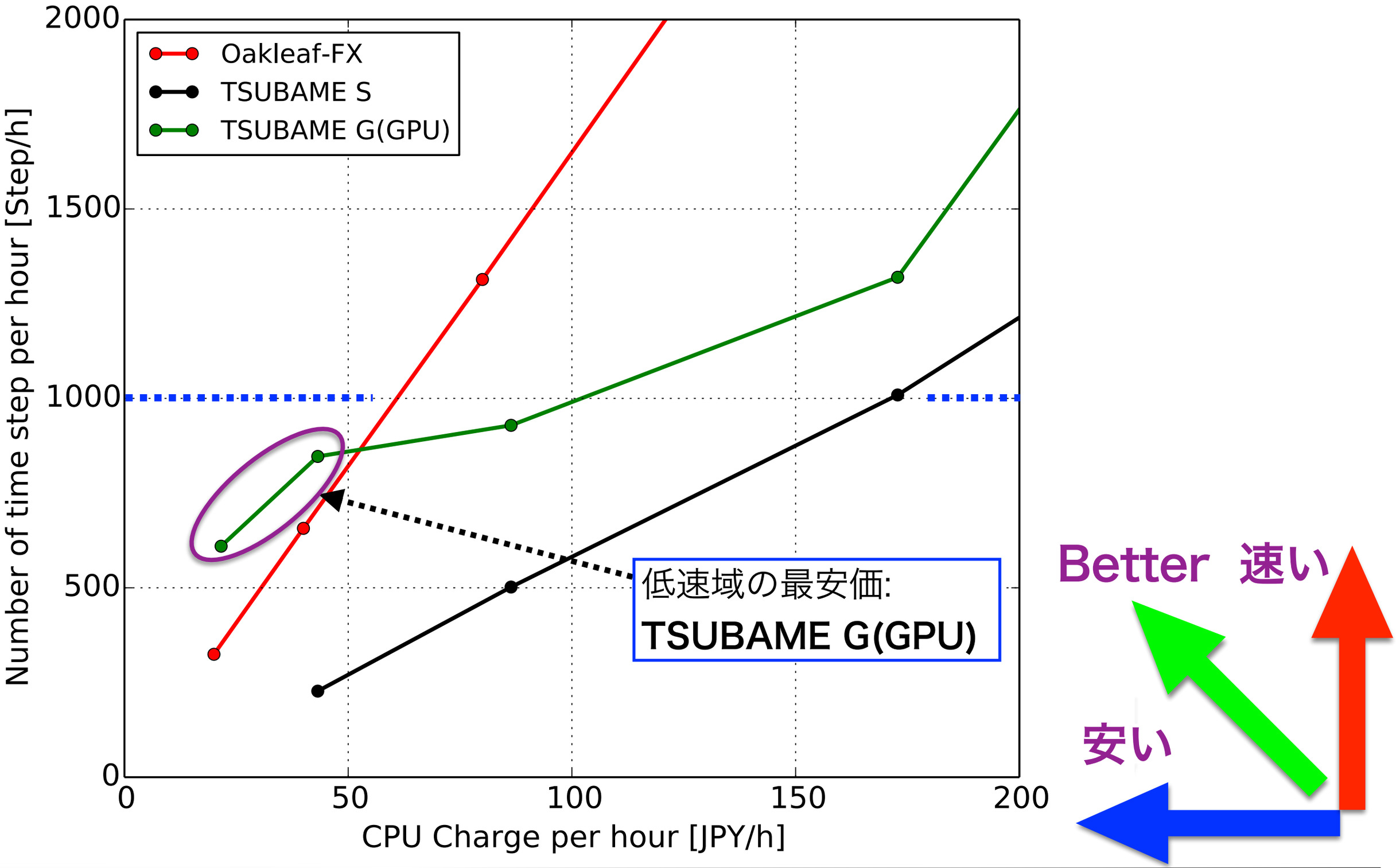
高速域での最安価:
TSUBAME S

FXでより多ノード数を使用した場合は不明
(FXのグループコースでは使用ノード数が申込ノード数を超えると課金が2倍になるので,申込ノード数に依存)



Copyright © 2015 OCAEL All rights reserved.

1時間の課金と時間ステップ数(成果公開型)



Copyright © 2015 OCAEL All rights reserved.

まとめ

- 格子数3Mのチャンネル流れを対象に，東京大学・東京工業大学・FOCUSのスパコン，Amazon EC2，Microsoft Azureのクラウドの各システムにおいて，OpenFOAMのベンチマークテストを実行した.
- 1時間で実行可能な時間ステップ数と課金額の関係から，時間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.
  - ✓ 成果非公開型(TSUBAME，FOCUS，EC2，Azure)
    - ✓ 高速域(6,000時間ステップ/h〜): FOCUS D
    - ✓ 中速域(1,000〜6,000時間ステップ/h): Azure, FOCUS D
    - ✓ 低速域(〜1,000時間ステップ/h): TSUBAME G(GPU)
  - ✓ 成果公開型(Oakleaf-FX，TSUBAME)
    - ✓ 高速域(6,000時間ステップ/h〜): TSUBAME S
    - ✓ 中速域(1,000〜6,000時間ステップ/h): Oakleaf-FX
    - ✓ 低速域(〜1,000時間ステップ/h): TSUBAME G(GPU)
- Amazon EC2のスポット利用は，料金が大きく変動するが，非成果公開型の低・中速で最安価になる場合も多いと予想されるので，今後詳細に検討したい.

Copyright © 2015 OCAEL All rights reserved.

謝辞

本研究では，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.

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.

ご静聴ありがとうございました

Copyright © 2015 OCAEL All rights reserved.