Vol.2020-HPC-174 No.2
2020/5/13

FOCUS スーパーコンピュータシステムにおける 並列課金インセンティブの効果 VII

西川武志[†1]

概要：これまで FOCUS スパコンの Intel Xeon 搭載システムは 2011 年 4 月共用開始の A システム（Westmere-EP）208 ノード，2014 年 1 月共用開始の D（Ivy-Bridge）80 ノード，2016 年 10 月共用開始の F，H（Broadwell-EP）前者 62 ノード，後者 136 ノード, 2018 年 7 月共用開始の V（Skylake）2 ノードの 4 つのアーキテクチャ，6 つの CPU 性能，メモリ容量，ネットワーク，3 種のアクセラレータ（Intel Xeon Phi, NVIDIA Tesla P100, NEC SX-Aurora TSUBASA），5 つのストレージ機器構成（単体 HDD，単体 SSD, RAID0 HDD, Lustre ファイルシステム, NFS ファイルシステム）の異なるシステムが存在する状況で課金インセンティブを適切に設定することで特定のシステムに大きく利用が偏ることが無いような運用を目指してきた．2011 年 4 月に稼働した A システムが現役で 9 年間運用され高い利用率を保っている一方で 2014 年 1 月導入の D,E システムの利用率が 2019 年度には急速に低下した．このような 4 世代にわたる計算資源が共存する中で主要な A, D, F, H のジョブ利用傾向について詳細な報告，並列度の推移，計算資源利用量，システム間の遷移などについて報告する．

キーワード：インセンティブ設計，計算センター運用，運用統計，並列度向上

1. はじめに

これまでの報告[1-6]では計算科学振興財団（FOCUS）が運用する産業界向けエントリースーパーコンピュータシステム「FOCUS スパコン」に関して 1 ジョブで複数のノードを確保して使うとノード課金時間単価を割引くという並列課金インセンティブを毎年度変化させて設定することで新規利用課題に対してはジョブ実行時の並列度を増加させることを図ってきた．

前回までの報告では図 1 に示すように更新回数が 4 回以上と多い利用者は経常的な計算資源のアウトソーシング先として FOCUS スパコンを利用し並列化を向上させる課金インセンティブに反応しない傾向がある．更新回数が少ない利用者に関して 2017 年度（マーカー内更新回数黒字）から 2018 年度（マーカー内各システム着色）への更新回数が 1 増えたものの変化に着目すると年度更新で並列化向上インセンティブに反応して平均並列数が増大する課題も存在し効果があることも判明した。

更新年度が少ない利用者で並列化向上インセンティブに反応しない利用者にヒアリングした結果，商用ソフトウェアを利用している利用者では，商用ソフトウェアの並列ライセンス数の制限の影響を大きく受けることやステップ状に課金単価を変化させていると不連続点で局所最適化することが判明した．

商用ソフトウェアの並列ライセンス数への柔軟な対応や課金単価の不連続点の影響を減少させるため 2019 年では図 2 に示す様に連続的に課金単価が変化する課金インセンティブを設定した．

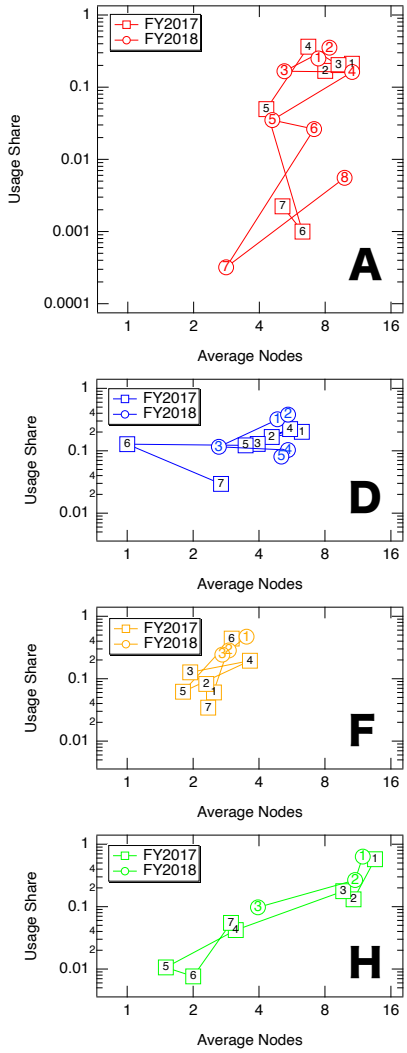

図 1　2017 年度および 2018 年度の利用者毎の更新年度数とA，D，F，Hシステムの平均並列ノード数に対する計算資源利用量全体に対する比（マーカー内更新回数）

†1(公財)計算科学振興財団
Foundation for Computational Science

ⓒ 2020 Information Processing Society of Japan

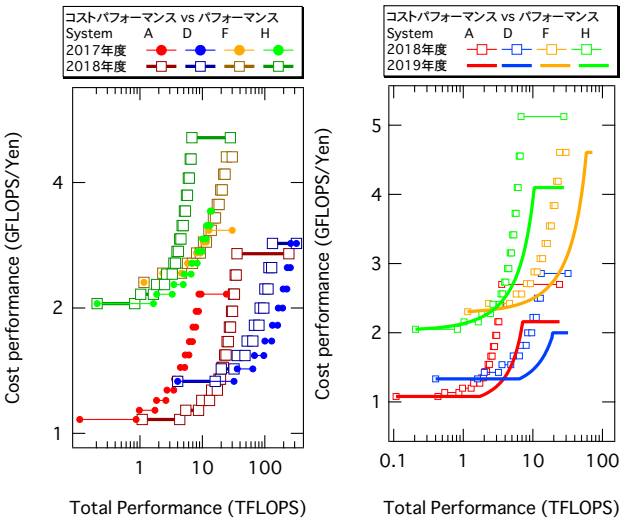

図 2　2017-2018 年度および 2018-2019 年度の課金インセンティブの変更状況

さらに 2017 年度から 2018 年度では F, H システムのコストパフォーマンスより A, D システムのコストパフォーマンスが A, D システムの合計性能が大きくなる領域，すなわち大規模ノード並列領域で上回る状況が存在したが 2018 年度から 2019 年度には全ての領域で A, D システムのコストパフォーマンスを F, H システムが上回る様に設定した．本稿では 2019 年度を通した利用実績を踏まえて個々の利用者をリピート回数に着目して年度毎の詳細な利用推移を調査し，課金インセンティブが個々の利用者にどのような利用傾向をもたらしているかを報告する．

2. FOCUS スパコンの各システムと課金インセンティブ

FOCUS スパコンシステムの概要についてはこれまでの報告[3,4,5]に A，D，E，F，H，V システムについて概要を述べているが，簡単に本項で今回の調査対象である A, D, F，H の各システムの基本仕様を述べる．

2.1 FOCUS スパコン A, D, F, H, V システム概要

FOCUS スパコン A，D，F，H システムの概要は次の通りである．

(1) A システム（208 ノード）

富士通ブレードサーバーPRIMERGY BX900 システム
高並列化環境（40Gbps QDR-IinfiniBand 接続）
CPU：Xeon L5640（Westmere-EP） 2.26 GHz 6 コア×2
108GFLOPS, RAM：48GB, HDD：500GB
JST からの委託事業により貸与され,後に譲渡を受け運用している FOCUS スパコン A システムは Westmere-EP 世代であり既に 2011 年の導入から 10 年度目を迎えた 2020 年度も運用を継続している．ただし 2018 年度から大規模分散ストレージシステム追加導入のため 2018 年度から 224 ノードを 1 ラック減らし 208 ノードに削減している．2011 年 3 月の導入時にランニングコストとしての電気代節約を目的に L5640 省電力 CPU を導入したため同時期のハイパフォーマンスを求めた X5xxx やコストパフォーマンスを求めた E5xxx が電力性能比で採算が合わなくなった状況とは対照的に現在も採算が取れた上で運用できている．

(2) D システム（80 ノード）

Cray CS300 システム
高並列化環境（56Gbps FDR-IinfiniBand 接続）
CPU：Xeon E5-2670 v2（Ivy-Bridge） 2.5 GHz 10 コア×2
400GFLOPS, RAM：64GB, HDD：6000GB
国からの補助金により Xeon Phi を 4 基搭載した E システム（故障ノードが 8 割を超えたため 2019 年度 48 ノード中 4 ノードを残して運用停止, 2020 年度からは全ノード運用廃止）と共に 2014 年 1 月に導入し 6 年度に渡って運用して来たが 2020 年度からは故障ノードが全体の 4 割を超えたため全ノード運用停止中である．

(3) F システム（62 ノード）

富士通 PC クラスタ PRIMERGY CX400 システム
高並列化環境（56Gbps FDR-IinfiniBand 接続）
CPU：Xeon E5-2698 v4（Broadwell） 2.2 GHz 20 コア×2
1152GFLOPS, RAM：128GB, HDD：6000GB
既存 A～D システムの運用収入の剰余を積み立てた資金で次項の H システムと同時期に 2016 年 10 月に導入し,現在の FOCUS スパコンの主たる計算資源である．

(4) H システム（136 ノード）

日本電気 Scalable Modular Server DX2000 システム
高密度高並列化環境（34 ノード/3U シャーシ，シャーシ間 40Gbps Ethernet ×16 シャーシ内ノード間 10Gbps Ethernet ×2 接続）
CPU：Xeon D-1541（Broadwell） 2.1 GHz 8 コア×1
205GFLOPS, RAM：64GB, SSD：512GB
組み込み向け CPU を搭載しデスクトップ仮想化, Web インフラストラクチャ DWH(データウェアハウス)，ビッグデータ解析用途に発売された製品を HPC 用途に, A システムの代替・後継として導入し F システムと同じく FOCUS スパコンの主たる計算資源である．

このように FOCUS スパコンシステムはノード性能と課金単価は 100 円/ノード時間の A, H システム 108GFLOPS から 500 円/ノード時間の F システム汎用 CPU1,152GFLOPS まで課金単価で 5 倍もの開きが存在する．利用者がどのシステムを利用すれば良いか実際のアプリケーションでコストパフォーマンスを見積もるため FOCUS ではシステムの理論性能のみならず複数の実アプリケーション（Advance/ParallelWave, ABINIT-MP, ANSYS Fluent, iconCFD, NAMD, OpenFOAM, STAR-CCM+）の性能についてベンチマーク結果を開示している[7].

2.2 FOCUS スパコン課金制度とインセンティブ

FOCUS スパコンの課金制度には大きく二つあり従量制と予約制がある[8]. 従量制は常設キューを利用したノード数と経過時間の積にノード時間単価を乗じて利用課金する. 予約制は事前に 1 日単位、月単位、年度単位でノード数を確保し, 1 日単位の予約は従量制同一の時間単価で課金インセンティブを与えて高並列利用を促進する意図を持って計算資源を提供している. 図 2 に示した様にA, D, F, Hシステムの課金インセンティブ（従量制時間単価）を毎年度変化させることで並列度を向上させる他, 他の大型計算センターや商用クラウドシステムが最新の CPU を採用し利用者に提供するのに対しコストパフォーマンスで見劣りしない計算資源提供も念頭においている.

2018 年度に対し 2019 年度の課金インセンティブの設計方針は以下の通りである.

- 全てのシステムでノード数増加に応じて連続に課金単価を下げた. ただし課金単価が 1 ノードの半額になったところで一定とした.
- A, D からの移行を促すため全てのノード数の領域で F, H のコストパフォーマンスを上回らせた.
- 故障の多い D から A への移行を促すため A のコストパフォーマンスは 36 ノード以上で D を上回らせた.
- 全てのシステムで並列インセンティブは前年度同一並列数では縮小し並列数の増加を促した.

3. 課金インセンティブの効果

各年度における各システムの常設 24 時間キューと日単位予約キューで利用された計算資源量：利用「計算資源比」, 2011 年度に A システムで利用された計算資源量を基準として各システムのノード時間を除して各システムの倍精度理論演算性能の A システムとの比を乗じた値と「平均ノード数」を 表 1 に A システム, 表 2 に D システム, 表 3 に F システム, 表 4 に H システムについて示す.

表 1 から A システムにおいて 24 時間キューと日単位予約キューの両方で平均ノード数が 2018 年度から 2019 年度に大幅に上昇したことがわかる. 利用計算資源量も 24 時間キューでは前年度とほぼ変わらず, 日単位予約キューでは 75%増となった. InfiniBand でノード間が結合された A システムは 10Gigabit Ethernet でノード間が結合された H システムよりも大規模並列ジョブが通信遅延や通信時間が小さくて済むため利用が手堅い.

表 2 からは D システムが 24 時間キューは利用が前年度比 44%と半減以下となり, 日単位予約キューは消滅同様となり, 平均ノード数も減少してしまった. この減少は D システムから他の A, F, H システムに移行したのではなく, 純粋に利用者が居なくなり, 利用も激減した影響であった.

表 1　2011 年度から 2019 年度における A システムの 24 時間キューと日単位予約キューで利用された計算資源量：利用計算資源比（FY2011 の A システムの利用量＝1）と平均ノード数

FY	24時間キュー		予約	
	計算資源比	平均ノード数	計算資源比	平均ノード数
2011	1.0	2.9	5.3	8.3
2012	2.6	3.8	2.6	3.8
2013	4.3	6.1	7.1	3.2
2014	1.2	7.2	4.3	1.8
2015	2.1	3.9	6.1	1.8
2016	1.5	2.5	1.8	1.3
2017	1.6	2.7	0.83	2.6
2018	1.0	1.7	0.14	11.8
2019	1.0	2.7	0.24	51.9

表 2　2011 年度から 2019 年度における D システムの 24 時間キューと日単位予約キューで利用された計算資源量：利用計算資源比（FY2011 の A システムの利用量＝1）と平均ノード数

FY	24時間キュー		予約	
	計算資源比	平均ノード数	計算資源比	平均ノード数
2014	3.2	2.8	1.7	3.1
2015	6.1	2.0	4.0	2.5
2016	4.0	3.3	1.3	7.7
2017	4.1	3.4	1.2	12.9
2018	3.6	2.5	0.60	2.5
2019	1.6	1.9	0.00	1.1

表 3　2011 年度から 2019 年度における F システムの 24 時間キューと日単位予約キューで利用された計算資源量：利用計算資源比（FY2011 の A システムの利用量＝1）と平均ノード数

FY	24時間キュー		予約	
	計算資源比	平均ノード数	計算資源比	平均ノード数
2016	2.2	2.5	0.35	2.3
2017	3.4	2.2	0.39	2.1
2018	7.5	2.4	0.25	1.6
2019	6.9	2.9	0.41	1.6

ⓒ 2020 Information Processing Society of Japan

表 4　2011 年度から 2019 年度における H システムの 24 時間キューと日単位予約キューで利用された計算資源量：利用計算資源比（FY2011 の A システムの利用量＝1）と平均ノード数

FY	24時間キュー		予約	
	計算資源比	平均ノード数	計算資源比	平均ノード数
2016	1.4	2.2	0.062	2.3
2017	1.6	2.2	0.070	2.1
2018	1.1	4.3	0.044	1.6
2019	0.85	1.9	0.073	1.6

表 3 から F システムでは 24 時間キューは利用が前年度比 8%減となったが，平均ノード数は 2 割上昇した．日単位予約キューでは計算資源利用が 6 割増となったが，平均並列ノード数は前年度と変わらなかった．主に商用ソフトウェアの利用のため並列数を増やすより，同一並列数で多数のジョブを実行する方がソフトウェア利用のコストパフォーマンスが良いためであるということであった．

表 4 から H システムでは 24 時間キューは利用が前年度比 23%減，平均ノード数はさらに 56%減となってしまった．これも既存の大口利用者が居なくなってしまった影響であり，そもそもの利用が無くなれば課金インセンティブの効果を計る以前の問題であった．

幸い，年度末に日単位予約キューを確保し 30 から 50 ノードのジョブを投入する大規模並列の利用者が現れたため予約キューの計算資源利用は 66%増となり平均ノード数も維持できた．

4. 考察と 2020 年度課金インセンティブの修正

図 3 に 2020 年度のA，D，F，Hシステムの課金インセンティブ（従量制ノード時間単価）を示すと共に 2018, 2019 年度の課金インセンティブを示す[8].

D システムに関しては 2019 年度の課金インセンティブが効いたかどうかの前段階，利用そのものが激減したため判別できなかった．H システムについては類似の状況であった．A, F システムに関しては概ね課金インセンティブは効果があった．したがって 2020 年度の課金はノード数増大に従って課金単価を連続で減少させること継続しつつ，他の大型計算センターや商用クラウドシステムの最新の CPU にコストパフォーマンスで見劣らない様に 1 ノード単価の見直し（前年度比 25%引き下げ）や並列ノード数が少ない領域で課金インセンティブが強く働く様に改訂した．

2020 年度の課金インセンティブの設計方針は以下の通りである．

- 全システムで 2019 年度より少ないノード数で課金単価が減少するように設定した．課金インセンティブが飽和するノード数は A システムが 66 から 13 ノード、F システムが 51 から 19 ノード, H システムが 66 ノードから 6 ノードと設定した．
- 経年劣化で故障が増えると想定される A システムでは 13 ノード以上で課金インセンティブを一定にし，大規模並列ジョブの実行を抑制する．
- H システムは富岳 48 コア/ノードを意識して 48 コア並列となる 6 ノード以上で課金単価が半額とした．

2019 年度に強い負のインセンティブを利用者に課し利用者のシステム離れを起こしてしまった反省と他の大型計算センターや商用クラウドを意識した大幅な課金インセンティブの変更を行った．

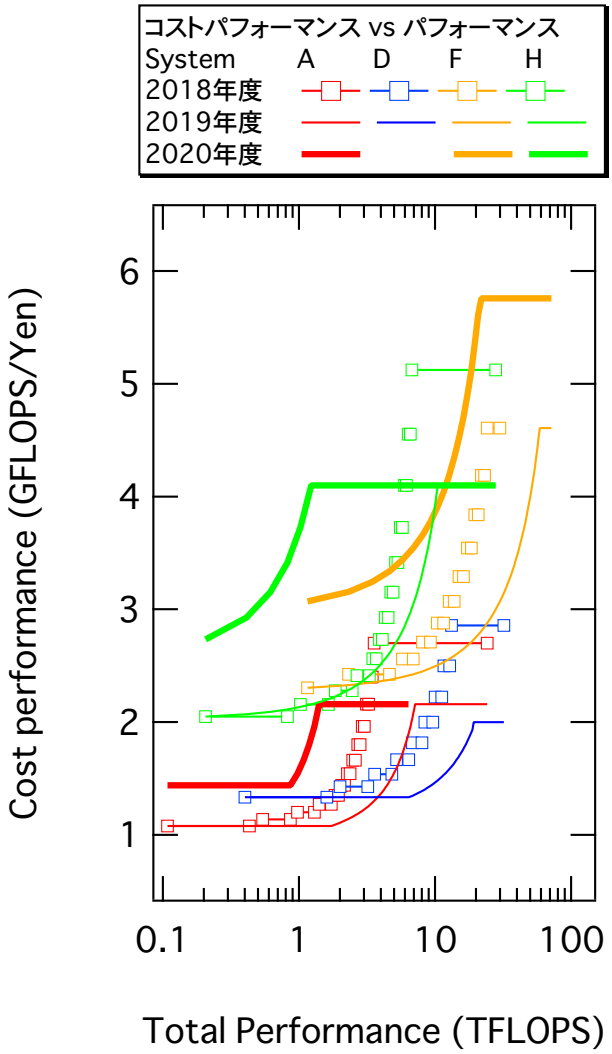

図 3　2018 年度から 2020 年度のA，D，F，Hシステムの課金インセンティブ（従量制ノード時間単価）

5. まとめ

本稿ではA，D，F，Hの複数のシステムの存在下で,ノード数増に応じて連続にノード課金単価を低減させる課金インセンティブを設計することでA,Fシステムでは並列度を向上させることに成功した一方で D, H システムでは2019 年度の並列度向上を強いた,前年度と同一ノード数では課金単価が下がらない課金インセンティブが有効に働かなかったことや大幅な外部環境変化（利用者が他の大型計算センターや商用クラウドに移行するのではなく，単純に利用が消滅）の前には多少の課金インセンティブでは太刀打ちできないことを改めて認識させられた.

一方で並列度の上昇を促すのは一筋縄では行かないことが明らかとなった.

過去の報告[6]で報告したFOCUSスパコンシステムに特化した並列度向上の講習会[9-11]は受講者が僅かである一方でAI/機械学習/ビックデータセミナー[12]は受講者が絶えない.

あらたな並列向上インセンティブとしてオンライン講習会や並列度の履歴に応じて課金単価を年度単位よりも短期で変動させる仕組みを現在検討中であるが，新型コロナウイルス感染拡大による政府の「緊急事態宣言」を受け、講習会の開催は中止となっており，講習会受講を活用した並列化促進にはオンラインセミナーの充実等で対処する予定である.

謝辞 FOCUS スーパーコンピュータシステムの運用や利用者の開拓に尽力されている計算科学振興財団の同僚と利用してくださっている利用者各位に，謹んで感謝の意を表する.

参考文献

[1] FOCUS スーパーコンピュータシステムにおける並列課金インセンティブの効果, 西川 武志, 研究報告ハイパフォーマンスコンピューティング（HPC）,2015-HPC-149(2),1-4 (2015-06-19).
[2] FOCUS スーパーコンピュータシステムにおける並列課金インセンティブの効果 II, 西川 武志, 研究報告ハイパフォーマンスコンピューティング（HPC）, 2016-HPC-157(10),1-5 (2016-12-14).
[3] FOCUS スーパーコンピュータシステムにおける並列課金インセンティブの効果 III, 西川 武志, 研究報告ハイパフォーマンスコンピューティング（HPC）, 2017-HPC-161(3),1-5 (2017-09-12).
[4] FOCUS スーパーコンピュータシステムにおける並列課金インセンティブの効果 IV, 西川 武志, 研究報告ハイパフォーマンスコンピューティング（HPC）, 2018-HPC-163(17),1-5 (2018-02-21).
[5] FOCUS スーパーコンピュータシステムにおける並列課金インセンティブの効果 V, 西川 武志, 研究報告ハイパフォーマンスコンピューティング（HPC）, 2018-HPC-166(1),1-4 (2018-09-20).
[6] FOCUS スーパーコンピュータシステムにおける並列課金インセンティブの効果 VI, 西川 武志, 研究報告ハイパフォーマンスコンピューティング（HPC）, 2019-HPC-170(23),1-4 (2019-07-17).
[7] FOCUS スパコンベンチマーク http://www.j-focus.jp/benchmark/
[8] FOCUS スーパーコンピュータシステム利用料金詳細, http://www.j-focus.or.jp/focus/fee.html
[9] FOCUS スパコン向け並列化講習会初級単体 CPU チューニング編, https://www.j-focus.jp/lectures/currently_accepting/focus-tuning2020a/
[10] FOCUS スパコン向け並列化講習会初級 OpenMP 編, https://www.j-focus.jp/lectures/currently_accepting/focus-tuning2020b/
[11] FOCUS スパコン向け並列化講習会初級単体 MPI 編, https://www.j-focus.jp/lectures/currently_accepting/focus-tuning2020c/
[12] AI/機械学習/ビックデータセミナー, https://www.j-focus.jp/lectures/currently_accepting/ai-ml-bd2020/

ⓒ 2020 Information Processing Society of Japan