PCクラスタにおけるMPIプログラムの仮想マシンへの割り当て手法

指導教員：
本多弘樹　教授
近藤正章　准教授
小宮常康　准教授

情報システム基盤学専攻
高性能コンピューティング学講座
本多研究室

西川　優

概要

近年、仮想化のHigh Performance Computing(HPC)分野での利用が注目されてきている。

長所：利用者は自分のニーズに合わせて物理コア数、
メモリ容量等をカスタマイズした仮想マシンを
利用できる。

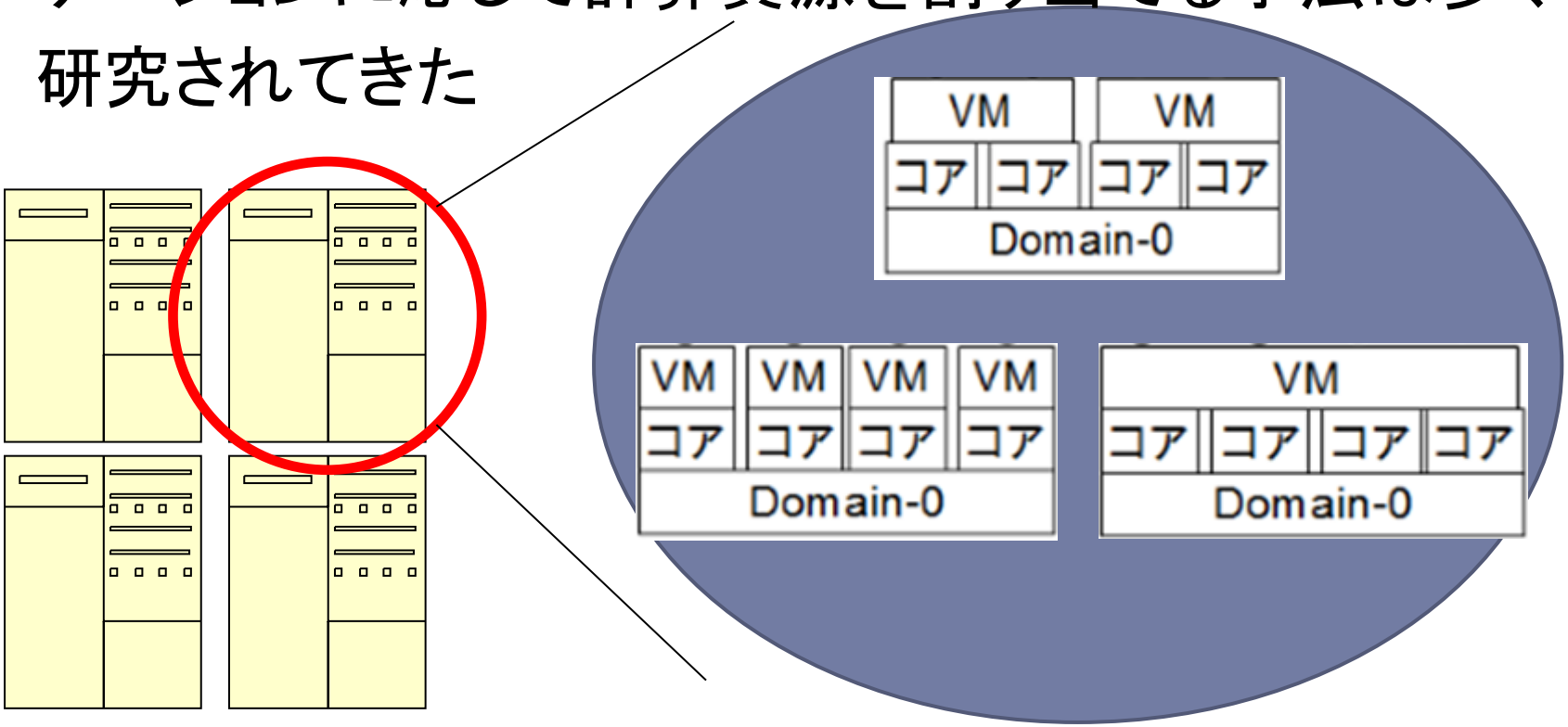
短所：パフォーマンスの低下。

企業の個人向けクラウドサービスのようなシステムを運用する際に重要な点として計算資源をどのように利用者に割り当てるかが挙げられる。

そこで本研究では仮想マシン環境においてMPIプログラムに応じてできるだけ実行時間が短くなる実行環境を選択する手法を提案していく。

関連研究[1][2]

これまでグリッド環境やPCクラスタ環境おいてアプリケーションに応じて計算資源を割り当てる手法は多く研究されてきた



[1]Fortaleza, Ceara, Brazil:"Supporting self-organization for hybrid grid resource scheduling"Symposium on Applied Computing Proceedings of the 2008 ACMsymposium on Applied computing Pages 1981-1986 Year of Publication: 2008
[2]長沼 翔、高橋 慧, 斎藤秀雄、柴田 剛志, 田浦 健次朗, 近山 隆:"ネットワークトポロジを考慮した効率的なバンド幅推定方法"先進的計算基盤システムシンポジウムSACSIS2008 pp359-366,2008

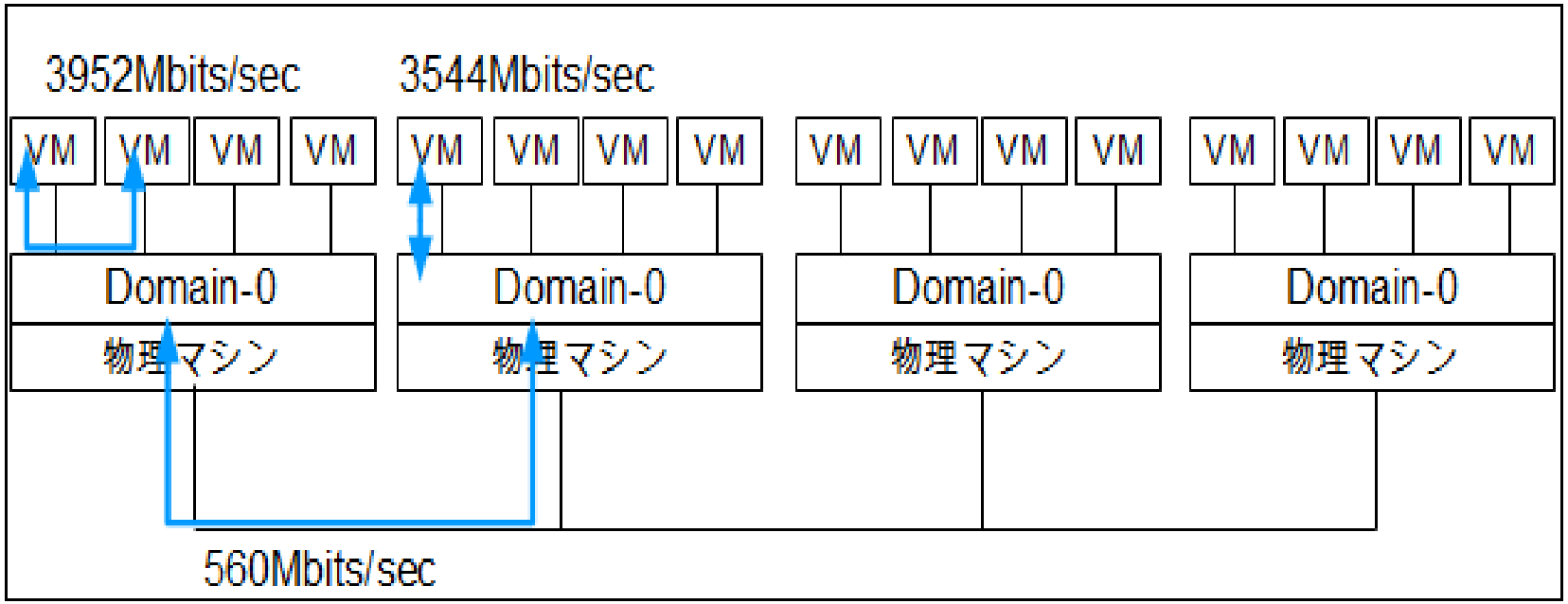
本研究の目的

本研究の目的は以下の3つが挙げられる。

- 物理マシン数、仮想マシン数、仮想マシンに割り当てる物理コア数の異なる実行環境でMPIプログラムを実行させ、実行環境と実行 時間の間に規則性があるかどうか解析する。
- 実行すべきMPIプログラムに応じてできるだけ実行時間が短くなる実行環境を選択する手法を提案する。
- 提案した手法の有効性を検証する。

これらの目的を満たすために予備実験を行った。
予備実験はいくつかの実行環境においてベンチマークプログラムを実行し、実行時間を測定した。

予備実験環境



memory	DDR3 1333MHz 4GB×3
CPU	GenuineIntel Core(TM)i7 3.07GHz
Kernel	2.6.18-194.11.3.el5xen
物理マシン数	4

Nas Parallel Benchmark[3]

本研究で使用するベンチマークプログラム

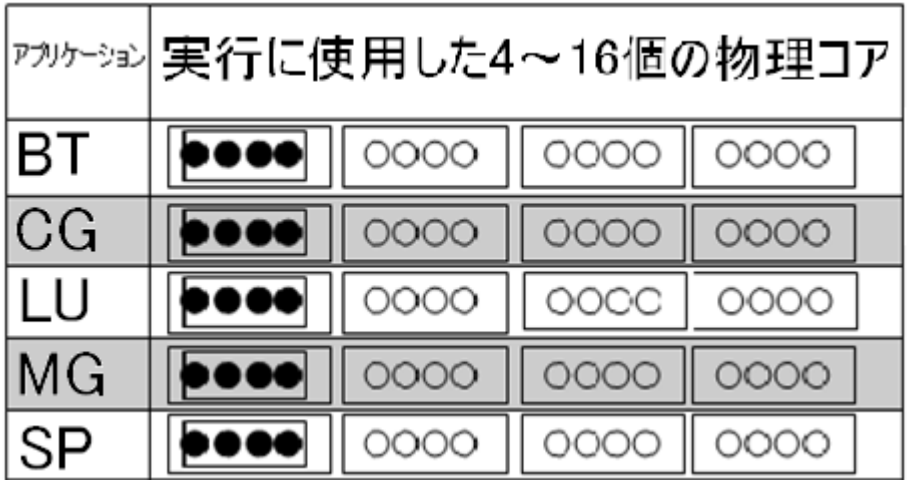
kernel		プロセス数
EP	乗算合同法による一様乱数、世紀乱数の生成	2の乗数
MG	簡略化されたマルチグリッド法のカーネル	2の乗数
CG	正値対称な大規模疎行列の最小固有地を求めるための共役勾配法	2の乗数
FT	FFTを用いた3次元編微分方程式の解法	2の乗数
IS	大規模整数ソート	2の乗数
Simulated CFD Application Benchmarks		
LU	Synmetric SOR iterationによるCFDアプリケーション	2の乗数
SP	Scalar ADI iterationによるCFDアプリケーション	nの2乗
BT	5*5 block size ADI iterationによるCFDアプリケーション	nの2乗

問題サイズ	class S	class W	class A	class B	class C	class D	class E
計算サイズ 繰り返し回数	小さい 少ない	⟷					大きい 多い

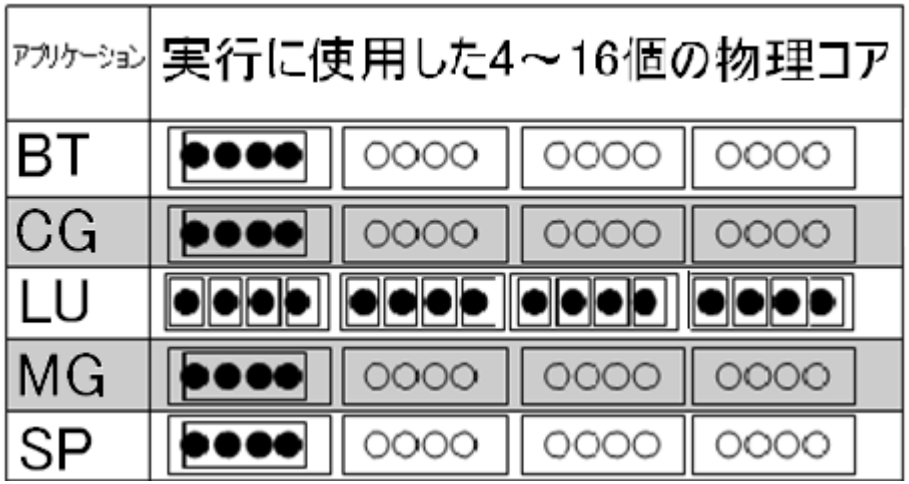
[3]Nas Parallel Benchmark :http://www.nas.nasa.gov/Resources/Software/npb.html

予備実験結果(仮想マシン環境)

いくつかの実行環境の中で実行時間が最も短かった実行環境を以下に示す



問題サイズ : Class B

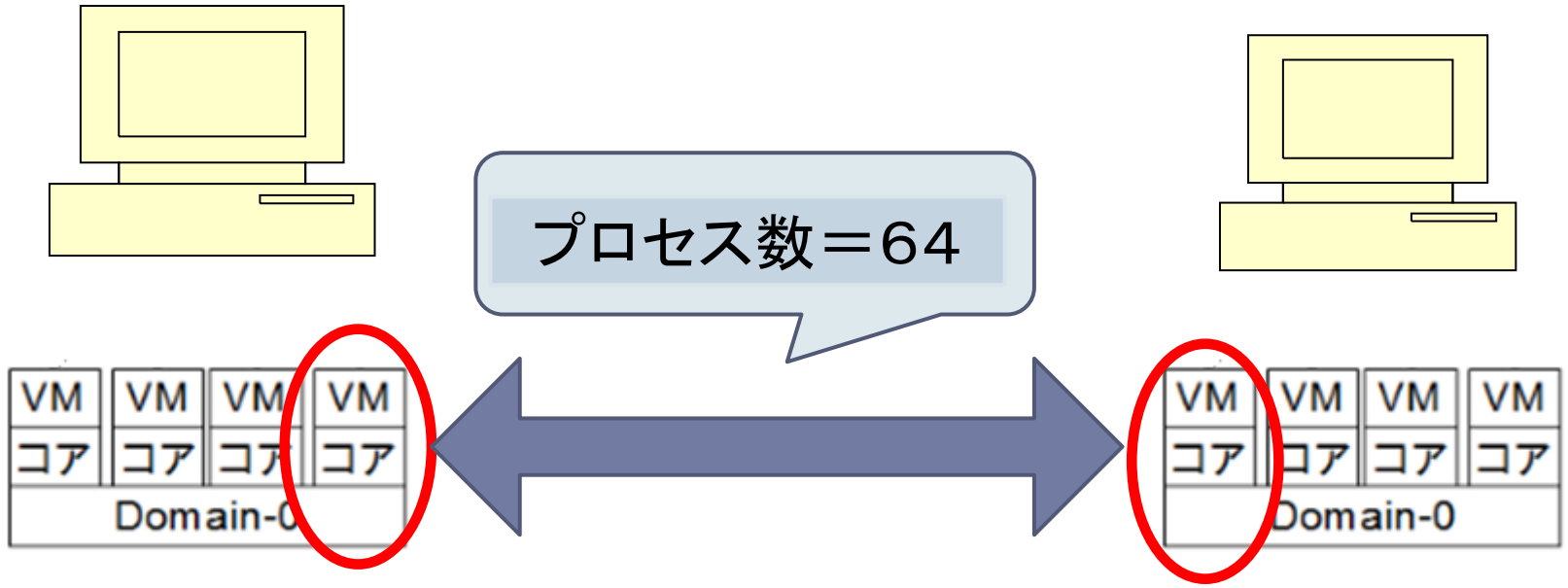


問題サイズ : Class C

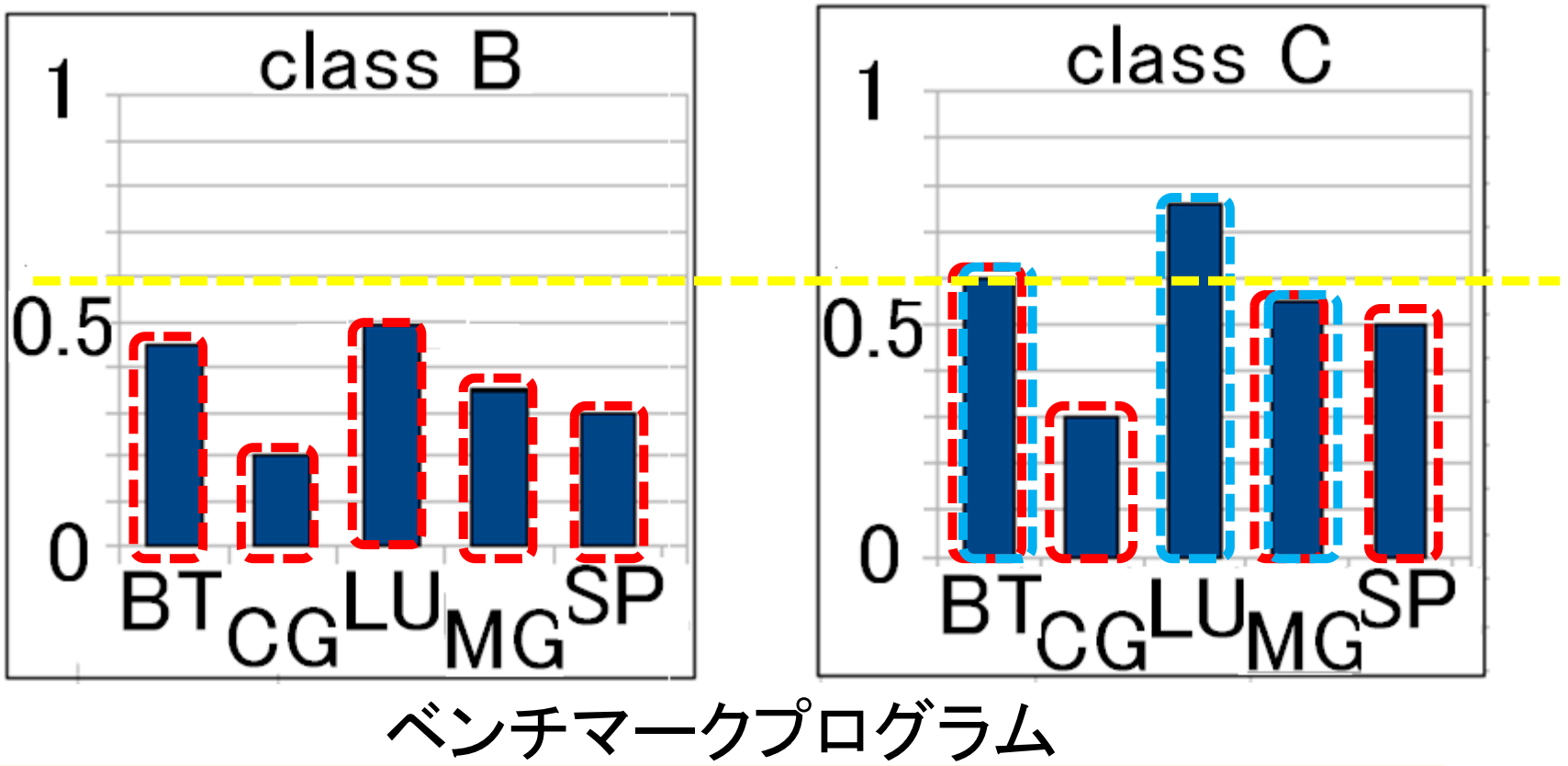
実行環境を選択する際に基準となる値の解析
CPU使用率の測定

CPU使用率＝単位時間あたりの特定の物理コアの使用時間の割合の最頻値

以下の実行環境でCPU使用率を測定する。



CPU使用率の測定


CPU使用率
class B
1
0.5
0
BT
CG
LU
MG
SP
class C
1
0.5
0
BT
CG
LU
MG
SP
ベンチマークプログラム
できるだけ実行時間が短くなる物理マシン数＝1
できるだけ実行時間が短くなる物理マシン数＝4
物理マシン数を決定する際の基準値

実行環境選択手法の提案

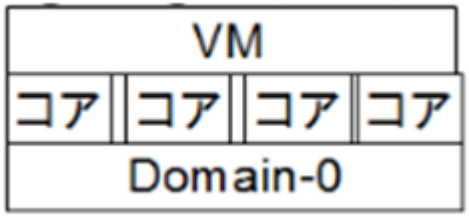
本研究で提案する実行環境選択手法の前提条件

- ユーザは実行すべきMPIプログラムと問題サイズを選択する。
- 実行に使用する物理マシン数、仮想マシン数、仮想マシンに割り当てる物理コア数はシステム運用者によって決められるものとする。
- MPIプログラムの実行に使用するプロセス数は実行に使用する物理コア数と同数とする。
- 実行すべきMPIプログラムのCPU使用率があらかじめ測定されているものとする。

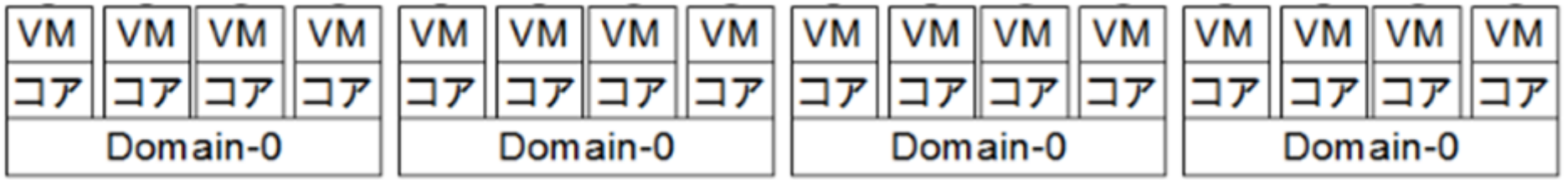
実行環境選択手法の提案

CPU使用率	$0 \leqq n \leqq 0.6$	$0.6 < n \leqq 1$
物理マシン数	1	4

実行に使用する物理マシン数が1の場合以下の環境において実行する。



実行に使用する物理マシン数が4の場合以下の環境において実行する。

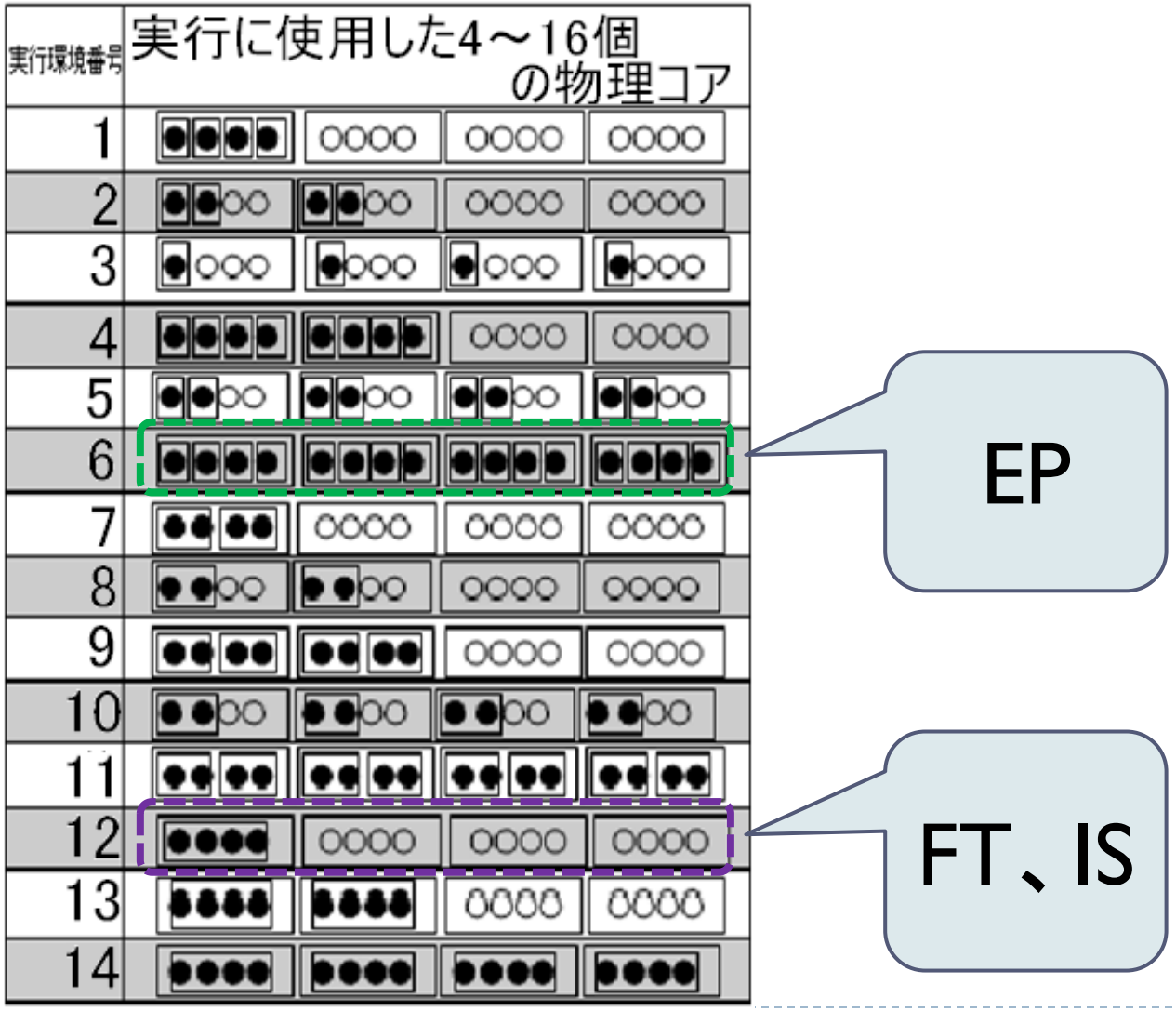


提案手法の有効性を示すための実験

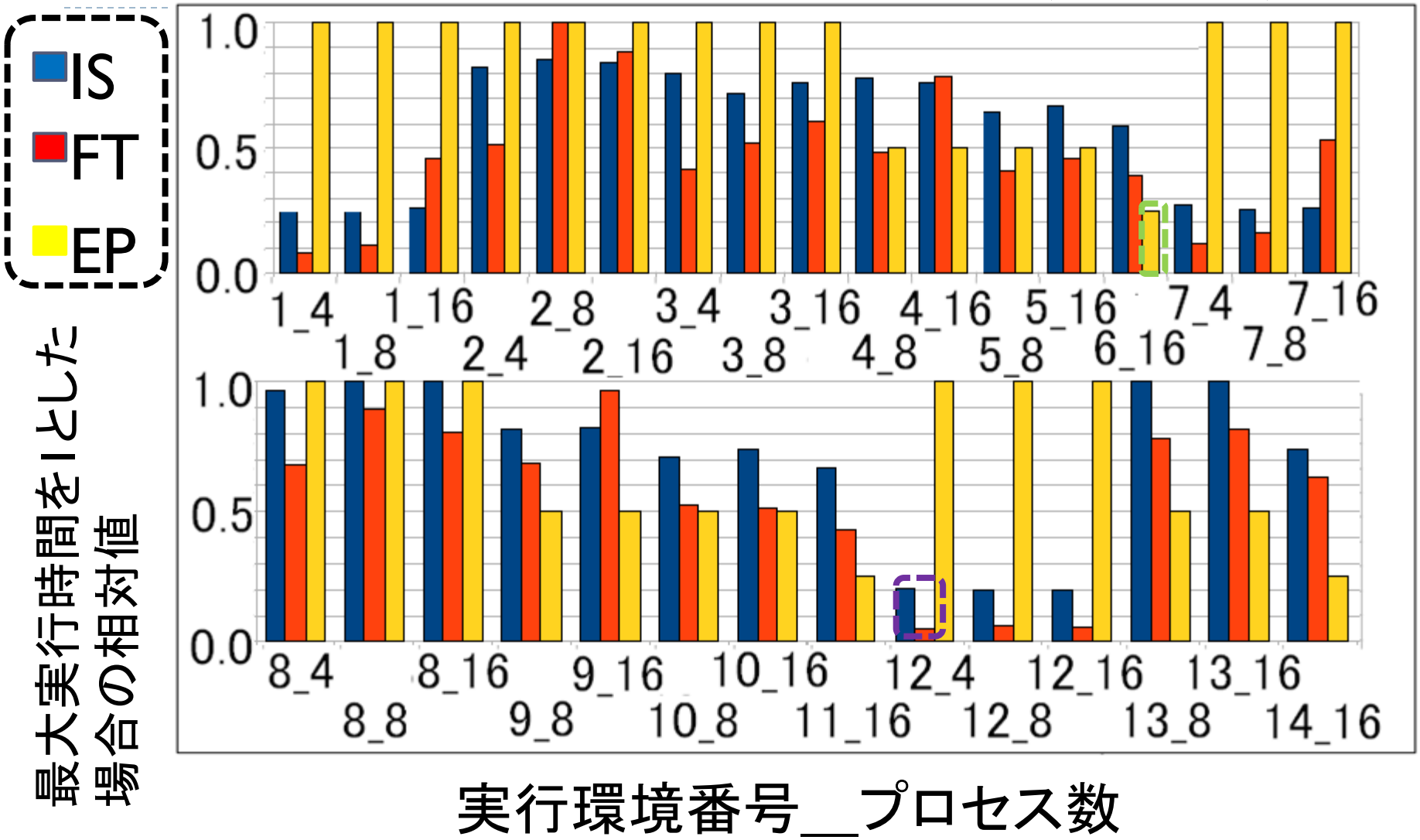
予備実験に使用していないIS、FT、EPのclass B、class Cに本手法を適用する。

MPIプログラム	問題サイズ	CPU使用率	選択される物理マシン数
IS	Class B	0.4	1
	Class C	0.4	1
FT	Class B	0.4	1
	Class C	0.5	1
EP	Class B	1	4
	Class C	1	4

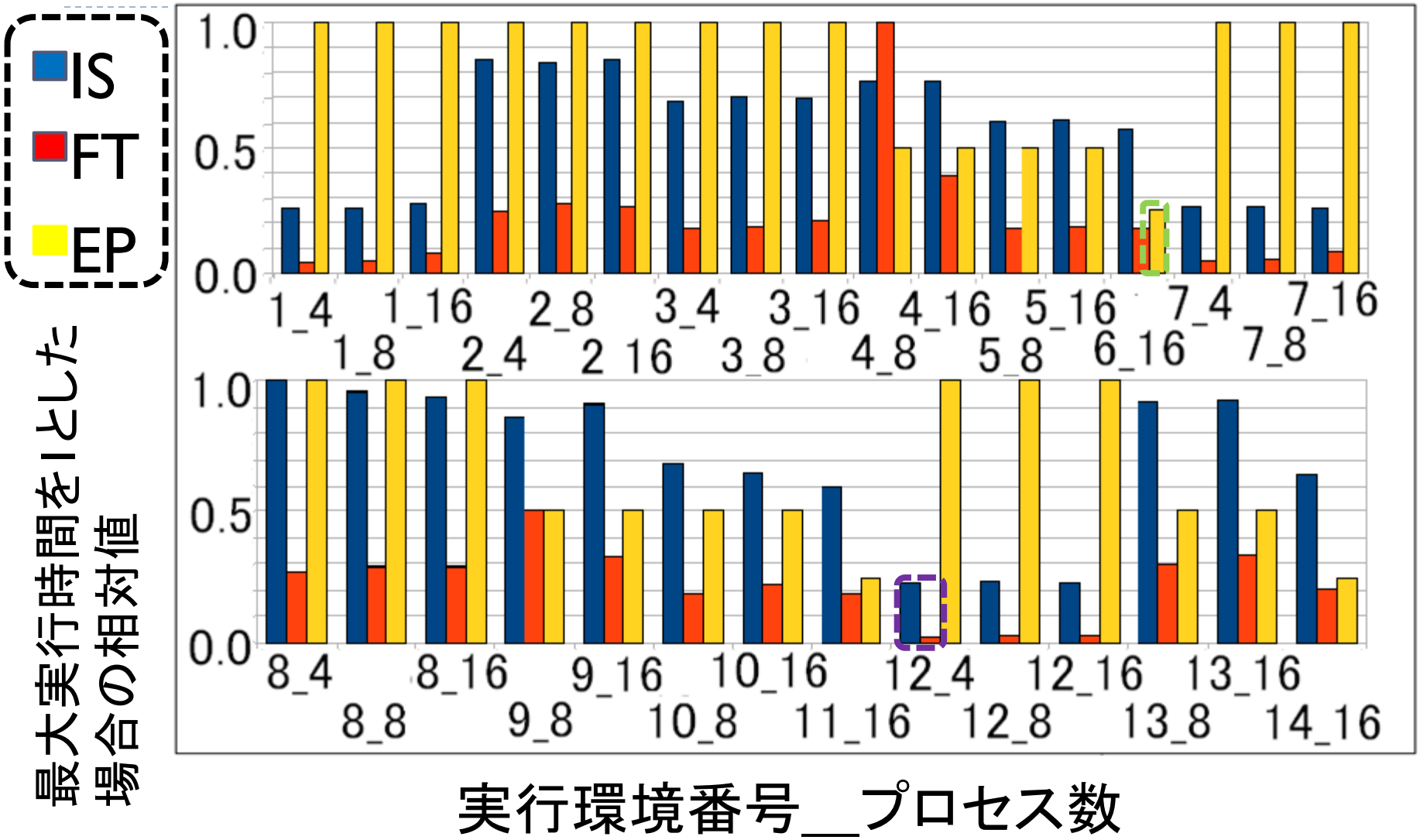
提案手法の有効性を示すための実験(class B)



提案手法の有効性を示すための実験(class B)



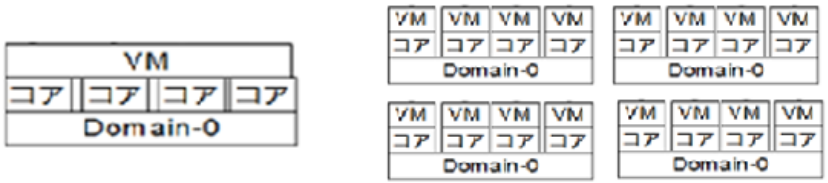
提案手法の有効性を示すための実験(class C)



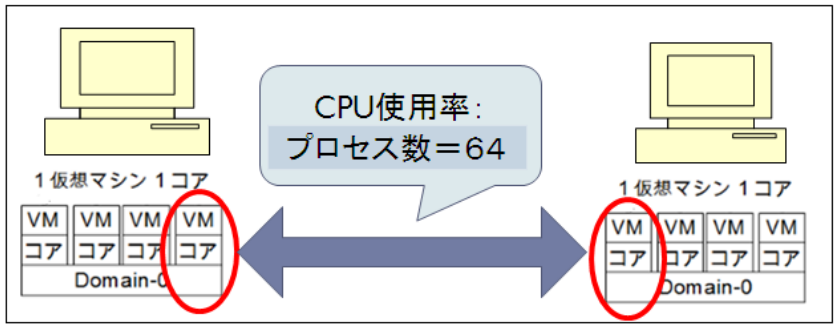
異なる実験環境において物理マシン数を選択する際に使用するCPU使用率の基準値を求める方法

いくつかのMPIプログラムを実行し、基準値を求める

- 実行時間を比較する



- 実行時間の差が誤差の範囲内と判断できる差となった場合、そのMPIプログラムを以下の環境で実行した場合のCPU使用率が基準値となる



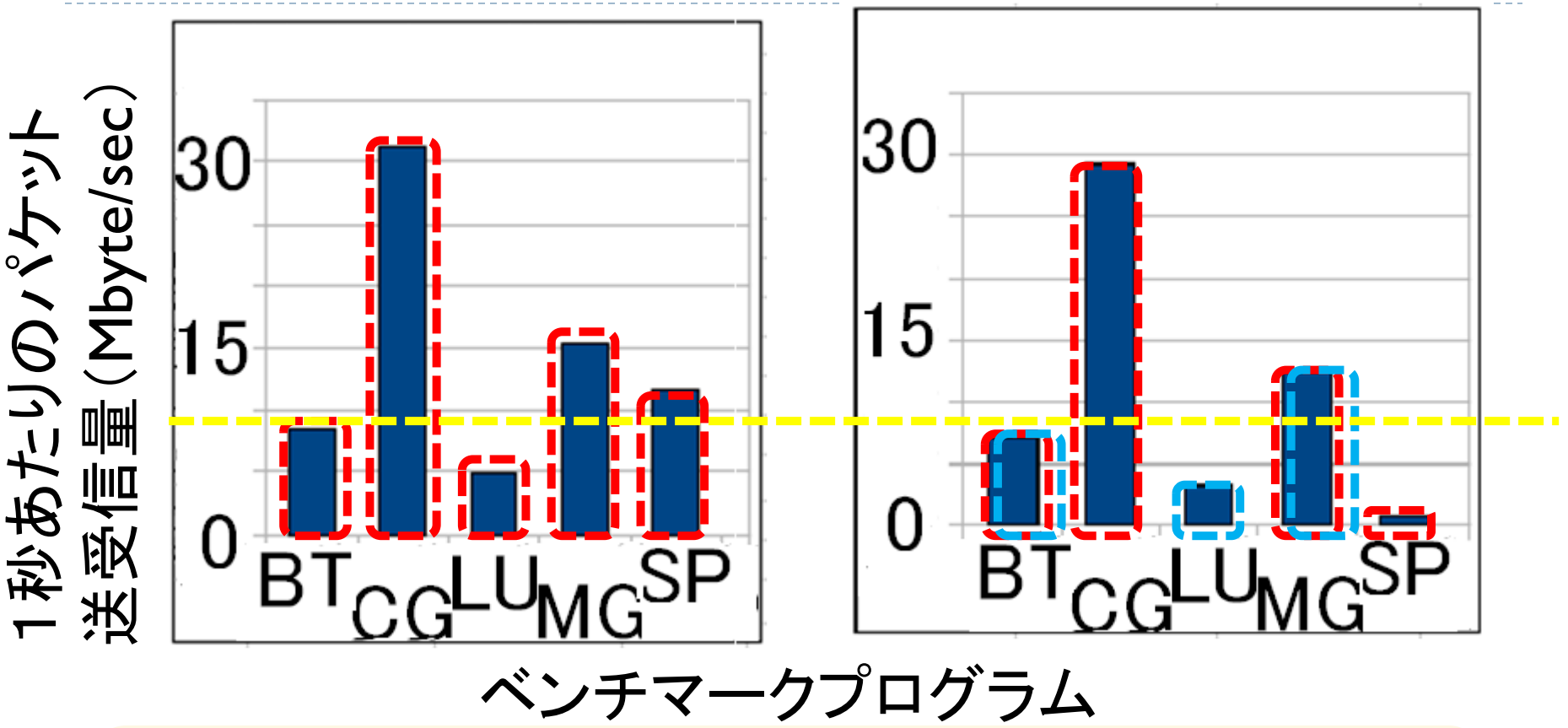
通信量を調節することでCPU使用率を自由に調節できるMPIプログラムを作成すると容易に基準値を調べることができる

今後の課題

今後の課題として以下の2点が挙げられる。

- 本研究の提案手法が実験環境の異なる場合においてもできるだけ実行時間が短くなる仮想マシン実行環境を提供できることを確認する。
- 物理マシン数を増やしていく場合に、できるだけ実行時間が短くなる実行環境がどのように変わっていくかを解析し、必要に応じて提案手法を改良していく。

付録：１秒当たりのパケット送受信量の測定



付録：予備実験結果(非仮想マシン環境)

できるだけ実行時間が短くなる実行環境を以下に示す

アプリケーション	実行に使用した4～16個の物理コア			
BT	●●●●	●●●●	●●●●	●●●●
CG	●●●●	○○○○	○○○○	○○○○
LU	●●●●	●●●●	○○○○	○○○○
MG	●●●●	●●●●	●●●●	●●●●
SP	●●●●	○○○○	○○○○	○○○○
EP	●●●●	●●●●	●●●●	●●●●

問題サイズ：Class B

アプリケーション	実行に使用した4～16個の物理コア			
BT	●●●●	●●●●	●●●●	●●●●
CG	●●●●	○○○○	○○○○	○○○○
LU	●●●●	●●●●	●●●●	●●●●
MG	●●●●	●●●●	●●●●	●●●●
SP	●●●●	●●●●	●●●●	●●●●
EP	●●●●	●●●●	●●●●	●●●●

問題サイズ：Class C