HITACHI
Inspire the Next

計算科学ロードマップからの計算機資源要件抽出

～特にI/O部分について～

2013/12/25　株式会社　日立製作所

本件連絡先：　masaaki.shimizu.sf@hitachi.com
toshiyuki.ukai.qq@hitachi.com

★ヒアリングシートの締め切りは2014/1/31で御願いします。
★ご質問、お問い合わせはお気軽にしてください。

HITACHI
Inspire the Next

Contents

1-1.おねがい

HITACHI
Inspire the Next

■目的

・エクサスケールシステムのI/O系設計に向けたアプリのI/O要件抽出

・そのために、本FSの皆様に各アプリのI/O要件のヒアリングに協力いただきたい

■現在までの準備状況

①I/Oヒアリングシート案作成(8/6に一度清水が説明)

②東大FSでのアプリ(RSDFT, COCO, ALPS, NICAM)のI/O要件入力でシート改良

③富田先生、西澤先生、八代先生協力で理研のアプリ(NICAM-LETKF)でシート改良

④本日、皆様にヒアリングシート入力をおねがいしたい

（本日説明、今月中にメーリングリストに展開します）

1-2.背景と目的

HITACHI
Inspire the Next

■目的
・エクサスケールシステムのI/O系設計に向けたアプリのI/O要件抽出

■課題
・アプリI/O要求パラメータの算出条件などの整理

■本資料の流れ
①エクサスケールシステムの構成と，アプリ実行条件などについて，現状開示できる範囲で一定の意識あわせ(スライド2, 3-1, 3-2)。
②I/O特性に応じてアプリを数種類に分類(スライド4)。簡易ヒアリングシートで基本パターンを提示いただいた上(スライド5)で，詳細ヒアリングシートによる整理(別紙)。
③基本パターンに収まらないアプリは個別ヒアリングを検討。

■考え方，留意事項
・アプリの規模感を合わせるため，エクサスケール全系でプログラムを実行することを想定。1ジョブで全系を使い切れない場合は，全系を埋めるようにジョブを多重化。このときの総I/O量を把握したい。
→全系を使う前提に合わない場合，簡易ヒアリングシート(スライド5)の特記事項にその旨ご記載ください。
・簡単化のために1プロセス/1ノードを想定。
・回答が困難な場合，例えば「問題規模」や「ジョブの実行時間」などは，京実行実績に基づき，「京におけるジョブ実行環境の〇倍」などの表現も可。
・記載にあたり，アプリのファイルIO振る舞い確認のため，京で動くI/Oプロファイルリングツール(ライブラリ)をご利用いただくことも可能(富士通様より情報提供いただいた)。

2. 想定するエクサスケールシステム

HITACHI
Inspire the Next

今回の前提：（総Flopsにあわせて全体が1/nになる可能性があります）

未確定の部分が多いが、ひとつの目安として下記数値を前提とした。

・総演算性能：1ExaFlops
・主記憶：10PB
・ノード数：10万ノード（　性能10TF/ノード、　メモリは50～100GB/ノード）

・高速ファイルシステム：100PB
　- 主記憶10PBの10倍

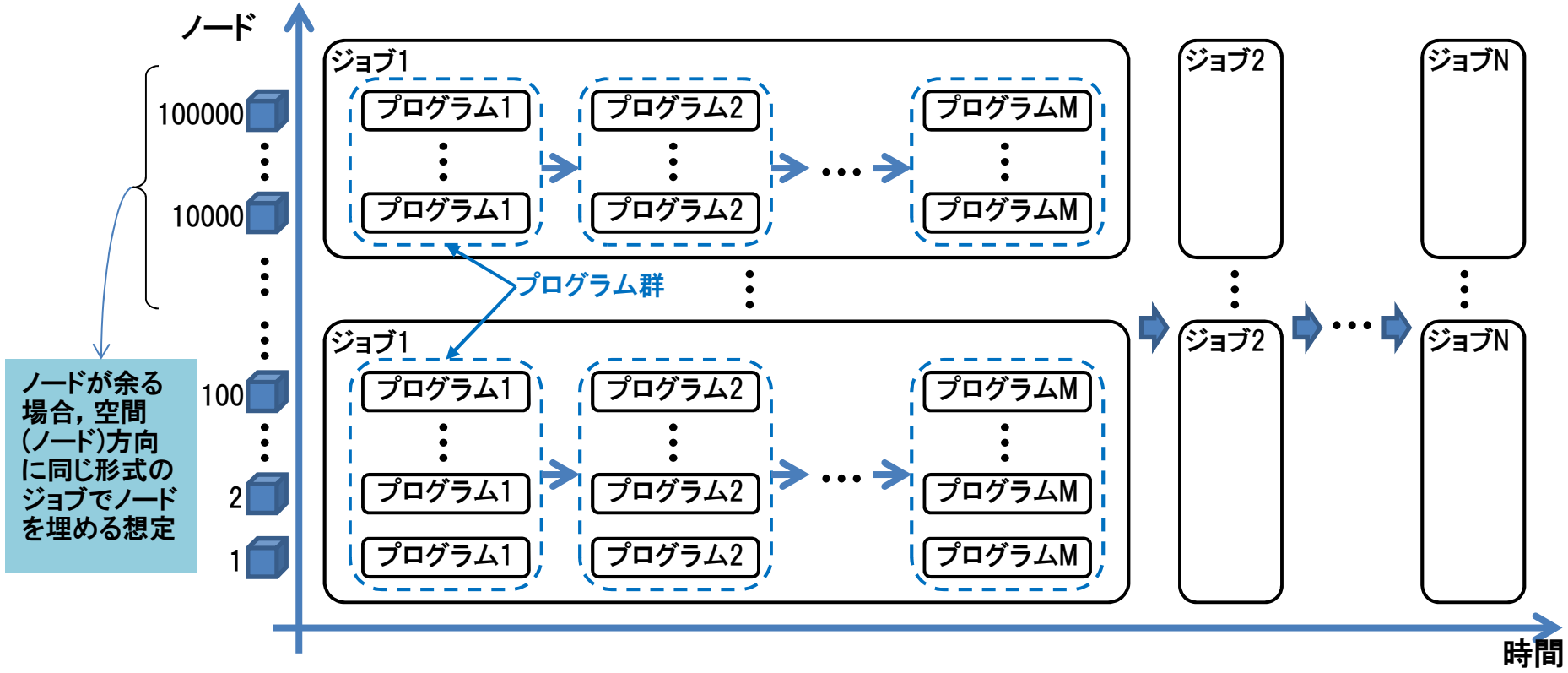
・高速ファイルシステム性能：10TB/s
　- 主記憶を1000秒で退避、アプリ実行時間の10%程度を目安

・総ファイルシステム容量：1ExaByte
　- 共有ファイルシステム、アーカイブを含んだ総量。

3-1. 用語定義 (ジョブ，プログラム)

HITACHI
Inspire the Next

■ジョブ関連の用語定義

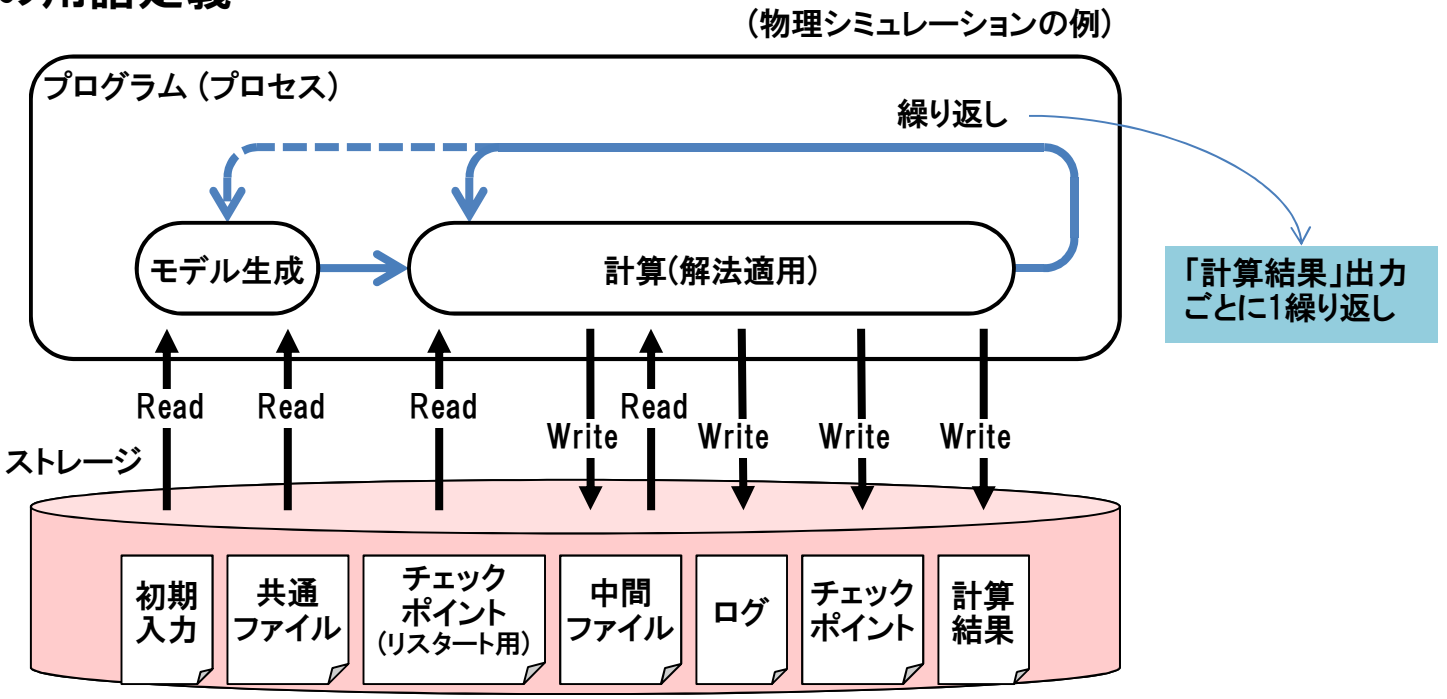


用語	説明
ジョブ	ジョブスケジューラに登録する「ジョブ」を想定。1ジョブは実行ファイルまたはシェルスクリプト。
プログラム	ジョブ実行時に実際に実行される実行ファイル。簡単化のために1プログラム(プロセス)/ノードを想定。京のステップジョブの場合のサブジョブ。
プログラム群	MPIで並列に動作するプログラム全体。

3-2. 用語定義(ファイル種別，繰り返し)

HITACHI
Inspire the Next

■ファイル関連の用語定義



#	用語	説明
1	初期入力	プログラム開始時にReadする初期データ。先行のプログラムの計算結果を引き継ぐ場合も含む。
2	共通ファイル	辞書などを想定。一つのプログラムが繰り返しRead，または，複数のプログラムで共有Read。
3	中間ファイル	作業用データ。プログラム内のみで利用。プログラム終了後は不要なデータ。
4	ログ	プログラムの実行状況，実行結果などの確認用。追記型。
5	計算結果	計算の結果，出力される解。プログラム終了後も必要なデータ。
6	チェックポイント	運用上の制限(使用時間制限を越える実行時間)，対障害，アプリ都合(一定ステップごとに出力など)で，ジョブ実行を再開できる情報。システムの信頼性が十分で，運用上の制限がない場合，I/Oする必要がないデータ。

4. アプリケーションのI/O特性による分類（案）

HITACHI
Inspire the Next

■狙い

・最終的にはアプリをI/O特性に応じて数種類に分類し，分類毎のI/O特性を押さえた上で要件を抽出。

■仮説

・アプリケーションによって、ファイルの利用目的による種別(初期入力，中間，計算結果，共通/共有の入力(辞書など))と，そのファイルのサイズや数，および，ファイルに対するプロセスのI/O特性などに基づき、数種類に分類できる。

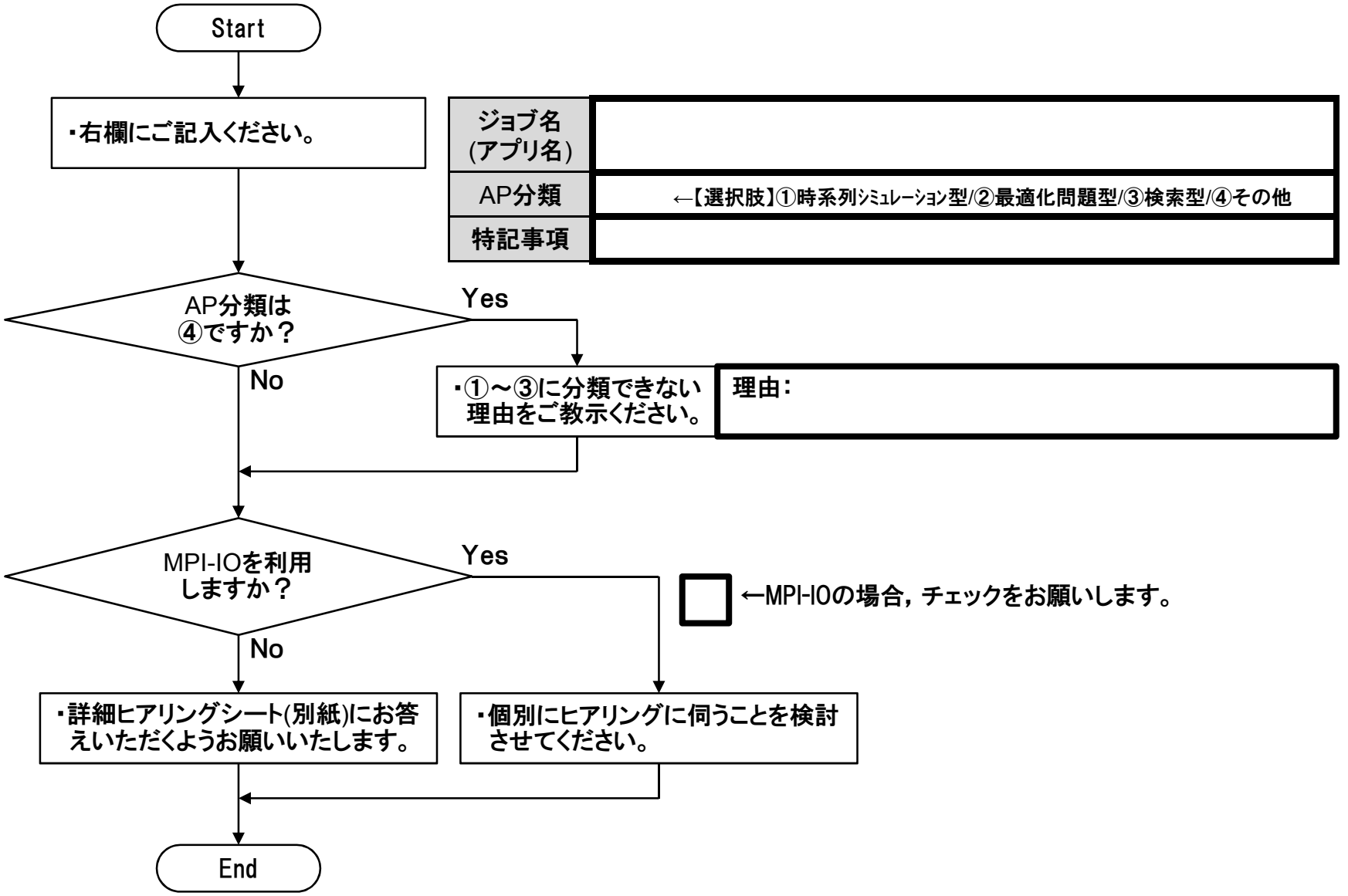
■分類案

・仮説に基づき，グランドチャレンジアプリのI/O分析結果を中心に検討。その結果，今のところ，①時系列シミュレーション型，②最適化問題型，③検索型の三種に分類。

#	分類観点	分類	トータルI/O容量			AP
			AP本体で必要なI/O		チェックポイント	
			初期入力	計算結果		
1	計算結果(時間発展する状態)を時系列に蓄積	①時系列シミュレーション型	小～大	大（上書き不可）	大（上書き可）	COCO, NICAM
2	最終の計算結果のみ必要	②最適化問題型	小～大	小～大（上書き可）＊再開のため最新数世代が必要な場合有	大（上書き可）	ALPS, RSDFT
3	入力ファイル数多	③検索型	中～大（ファイル数多）	小	N/A	ゲノムマッチング

小：GBクラス，中：TBクラス，大：PBクラス

5. 簡易ヒアリングシート



6-1. 詳細ヒアリングシートのサンプル (1/2)

HITACHI
Inspire the Next

◆別紙として詳細ヒアリングシートを用意。下記は，ジョブ実行条件など，全体概要を記載いただきたい箇所。

■お願い

・別紙説明書(計算科学ロードマップからの計算機資源要件抽出 ～特にI/O部分について～)を参照してください。
・ジョブはSMP MPIで，1プログラム(プロセス)/1ノードと想定してください。
・「全体概要」以外は，ジョブまたはプログラムがI/Oする個々のファイル単位でお答えください。
I/O時間が問題となる可能性がある，主要なファイルのみで結構です。
・エクサスケール世代のマシンでジョブ/プログラムを実行することを想定してください。お答えが難しい場合，例えば「問題規模」や「ジョブの実行時間」などは，京実行実績に基づき，「京におけるジョブ実行環境の〇倍想定」などの表現でも結構です。
・ご回答欄を全て埋めていただけなくてもご返答ください。「回答期待度」をご考慮の上，ご回答いただければ幸いです(◎：特大，〇：大)。
・ファイルの記載欄が不足する場合はフォームをコピーして追加してください。

ジョブ名	ジョブA
プログラム名	-
ご回答年月日	2014年 x 月 x 日

■全体概要

#	回答期待度	項目		回答
1	◎	実行条件(エクサスケール世代想定)	システム全体での同時実行ジョブ(プログラム群)数	2
2	◎		1プログラム群当たりの同時実行プログラム数(=MPI並列数=ノード数)	40000
3	〇		問題規模	160G Grid (40000x40000x100)
4			その他	時間レンジ 10年間をシミュレート，dt = 25s (約3,500 step / day)
5	〇	ジョブの実行時間		-
6	〇	期待するI/O性能(or 計算時間に対するI/O時間の割合)		10%
7	〇	プログラミング言語		
8		特記事項		【例】アプリ都合でのチェックポイント有無 etc.

6-2. 詳細ヒアリングシートのサンプル (2/2)

HITACHI
Inspire the Next

◆下記はI/Oに関してご記載いただきたい箇所。ファイル毎のご記載をお願いしたい。
(I/O時間が問題となる可能性がある，主要なファイルのみで結構です)

■ファイル					
1	#	回答期待度	項目		回答
	1	◎	ファイル名		INITIAL
	2	◎	格納データ概要		初期条件
	3	◎	I/O種別		初期入力
	4	◎	アクセス方法		(D) I/Oマスタ 代表プログラムでreadしてscatter (MPI-IO準拠の方式に移行予定)
	5	○	アクセスパターン(シーケンシャル，ランダム)		
	6	◎	I/Oするプロセスの数 (/プログラム群)		1
	7	◎	1ファイルの容量		12TB
				計算式 or 根拠	-
	8	○	1繰り返し当たりのファイル数		1
				計算式 or 根拠	-
	9	○	ファイルI/Oの繰り返し数 (/プログラム)		1回
				計算式 or 根拠	-
				上書き有無	N/A
	10	◎	総I/O容量 (/プログラム群)		12TB
				計算式 or 根拠	12TB x 1回
	11	◎	【writeの場合】 最終的に保持必要なファイル容量(/プログラム群)		N/A
				計算式 or 根拠	-
	12		1ファイルに対する1回当たりのI/O量 (I/O単位)		-
	13		ファイルのI/O頻度		初期設定時に一度のみ
	14		特記事項		アクセス方法はMPI-IO準拠の方式に移行予定

6-3. サンプル「ジョブA」 ①時系列シミュレーション型

HITACHI
Inspire the Next

実行条件	・システム全体で 2 ジョブを独立に同時実行 ・1ジョブあたり 200x200 = 40,000 プロセス ・ 問題規模　40,000 x 40,000 x 100　（160G grid） ・ 時間　10年間をシミュレート , dt = 25s　（約3,500 step / day ）
初期入力	・INITIAL（初期条件）ファイル ・ 容量 12TB を1回のbinary read ですべて読み込む ・ 現在は代表プロセスでreadしてscatter、MPI-IO準拠の方式へ移行予定
出力	・ 約10種類のモデル変数の値を、変数ごと1日（3,500step）ごとに出力 ・ 1ファイルサイズは 160G grid x 8B = 1.28TB、 1回に12.8TB （10変数）、全体で、12.8TB x 3,650日分 = 46.7PB ・ 現在は代表プロセスがgatherしてwrite、MPI-IO準拠の方式へ移行予定
チェックポイント	・ 1日分のシミュレート（3,500 step）ごとに出力 ・ ファイルは1つ、容量約12.8TB ・ 現在は代表プロセスがgatherしてwrite、MPI-IO準拠の方式へ移行予定

HITACHI
Inspire the Next

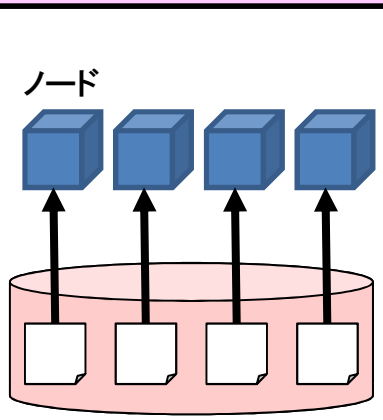
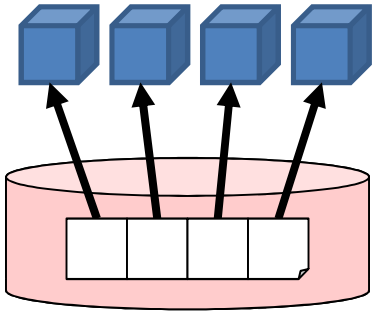
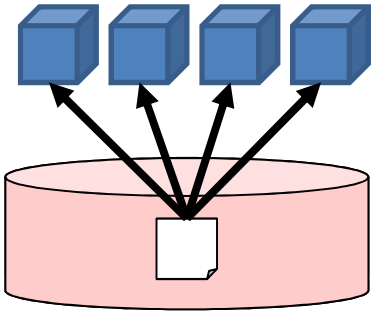
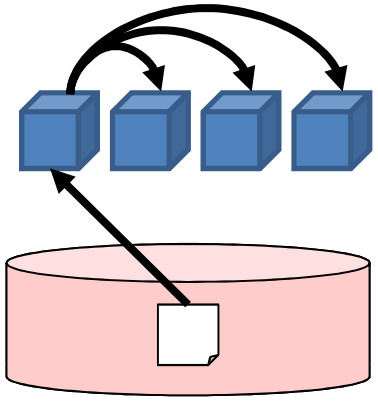
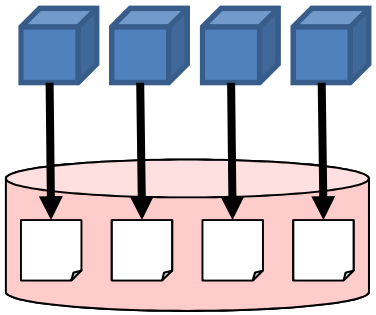
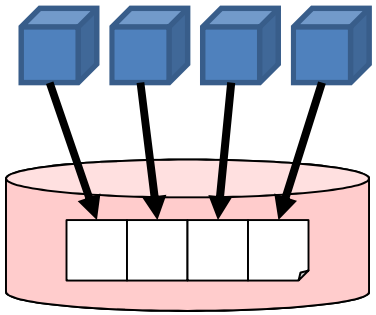
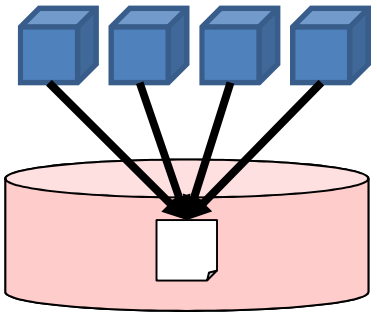
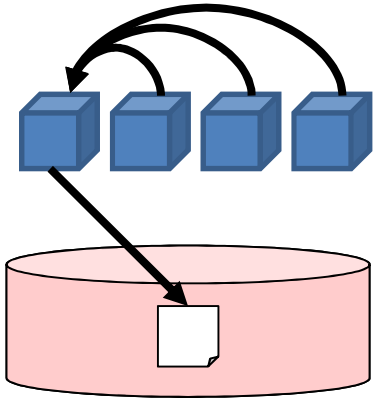
以下，付録

付録1. ファイルアクセスパターンの分類

HITACHI
Inspire the Next

■ファイルアクセスパターンの分類

★詳細ヒアリングシート回答時ご参考

分類	(A) 分散ファイル	(B) 共有ファイル1	(C) 共有ファイル2	(D) I/Oマスタ
Readするファイルのアクセスパターン	 各ノードが別ファイルをRead	 各ノードが同じファイルの別領域をRead	 全ノードが同じファイルをRead(初期入力，辞書等)	 マスタがファイルをRead
Writeするファイルのアクセスパターン	 各ノードが別ファイルにWrite	 各ノードが同じファイルの別領域にWrite	 全ノードが同じファイルに繰り返し追記(ログなど)	 マスタがファイルにWrite

付録2. サンプル「ジョブB」 ②最適化問題型

HITACHI
Inspire the Next

実行条件	・システム全体で 1000 ジョブを独立に同時実行 ・1ジョブあたり 100 プロセス ・ジョブの実行時間は24時間 ・チェックポイント用のデータを1時間毎に出力 ・24時間後にジョブは正常終了し、チェックポイントデータは削除されるとする
初期入力	・ プロセス0が、params.in.xml （1KB）, params.task1.in.xml （10KB） の 2ファイル（テキスト・XML）をプログラム開始時に1回だけ読む
チェックポイント / 出力	・「1回/時間 + ジョブ終了時」にプロセス0が出力、 （1） params.out.xml （テキスト・XML, 1KB） （2） params.task1.out.xml （テキスト・XML, 10KB） ・「1回/時間 + ジョブ終了時」に全プロセスが独立に出力 （3） params.task1.clone1.workerXX.h5 （バイナリ・HDF5） プロセス0 は1GB、他プロセスは各々10MB ジョブあたり 2GB （上書きしてよいが最新2世代分は残す）
チェックポイント	・ params.task1.clone1.workerXX （バイナリ, XDR） ・「1回/時間 」全プロセスが独立に出力 ・プロセスあたり 2GB, ジョブあたり 200GB

付録3-1. ご回答活用例 -既判明分のアプリI/O要件のまとめ-

HITACHI
Inspire the Next

◆ご回答は，下のように整理し，最低限必要な容量，I/O帯域の算出，その他のニーズの把握に利用させていただく。

				COCO大規模	ALPS小規模	ALPS大規模	RSDFT大規模	RSDFT中規模	NICAM-LETKF NICAMサイド	NICAM大規模
基本諸元	ジョブ数									
	システム全体	A		2	1,000	8	1	8	100	2
	プロセス数									
	ジョブ当たり	B		147,456	384	49,152	387,072	49,152	1,024	163,840
	システム全体	C		294,912	384,000	393,216	387,072	393,216	102,400	327,680
	ノード数(※)									
	ジョブ当たり	D		36,864	96	12,288	96,768	12,288	1,024	40,960
	システム全体	E		73,728	96,000	98,304	96,768	98,304	102,400	81,920
	問題規模									
	シミュレーション規模を特徴づける諸元	F		38,400x38,400x200 (295G格子点)			322,560原子、格子数=480x512x672、固有ベクトル数:663,552	110,592原子、格子数=384x384x384、固有ベクトル数:245,760	g-level=12, r-level=7, v-layer=100 (1765G格子点)	g-level=14, v-layer=94 (258G格子点)
	時間(シミュレーション時間)									
		G		10年間 (dt=25秒/1日=3,456step)	(実時間で24時間)	(実時間で24時間)	?	?	1ヶ月分 (dt=2秒/1時間=1,800step)	1ヶ月分 (dt=2秒/1時間=1,800step)
ストレージ諸元	入力(システム全体)									
	ファイル数	H		4	2,000	16	1	8	710,400	21,954,560
	入力量(GB)	J		94,400	0.011	0.000	0.0077	0.0027	269,275	105,022
	出力(計算結果、システム全体)									
	ファイル数/出力間隔	K		20	386,000	393,232	2	16	6,758,400	5,242,880
	出力量(GB)/出力間隔	L		47,200	4,840	3,940	876,700	890,400	976,451	67,200
	出力間隔(演算時間(s))	M		750	3,600	3,600	43,200	43,200	12,000	2,000
	延べ出力ファイル数	N		73,000	9,264,000	9,437,568	?	?	?	3,774,873,600
	延べ出力量(GB)	P		2,294,720	116,160	94,564	?	?	?	48,384,000
	保存ファイル数	Q		73,000	772,000	786,464	2	16	6,758,400	3,774,873,600
	保存ファイル量(GB)	R		172,280,000	9,680	7,880	876,700	890,400	976,451	48,384,000
	チェックポイント(リスタート用のデータ)									
	ファイル数/出力間隔	S		2	384,000	393,216			1,843,200	?
	出力量(GB)/出力間隔	T		43,000	768,000	786,432			245,740	?
帯域	出力のみの帯域									
	出力間隔の10%(s)	U	0.1xM	75	360	360	4,320	4,320	1,200	200
	必要帯域(GB/s)	V	L/U	629	13	11	203	206	814	336
	操作対象ファイル/s	W	K/U	0.27	1,072.22	1,092.31	0.0005	0.0037	5,632	26,214
	チェックポイントのみの帯域									
	出力間隔の10%(s)	α	0.1xM	75	360	360	4,320	4,320	1,200	200
	必要帯域(GB/s)	β	T/α	573	2,133	2,185	0	0	205	?
	操作対象ファイル/s	γ	S/α	0.03	1,066.67	1,092.27	0.00	0.00	1,536.00	?
	出力+チェックポイントの帯域									
	出力間隔の10%(s)	X	0.1xM	75	360	360	4,320	4,320	1,200	?
	必要帯域(GB/s)	Y	(L+T)/U	1,203	2,147	2,195	203	206	1,018	?
	操作対象ファイル/s	Z	(K+S)/U	0.29	2,138.89	2,184.58	0.00	0.00		?

172.3 PB

0.79 PB

814 GB/s

2,185 GB/s

2,195 GB/s

帯域の計算はI/O時間を演算時間の10%と仮定して実施

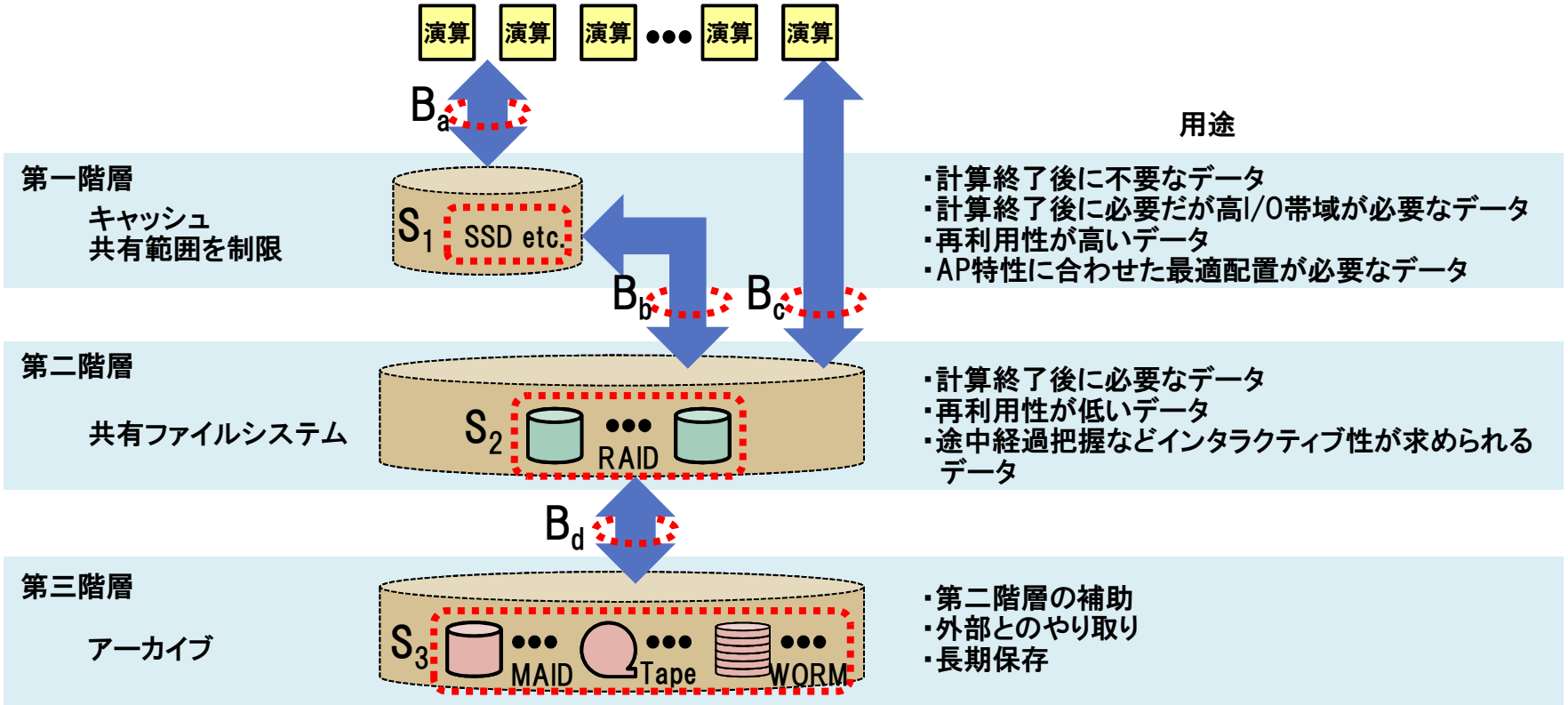
時系列sim.型　　最適化問題型　　時系列sim.型

付録3-2. ご回答活用例 -階層化ストレージの構成モデル-

HITACHI
Inspire the Next

◆要件　(a) 容量/帯域的に支配的な，初期入力，結果出力，チェックポイント出力をカバー。
(b) 共用利用するデータ(数万ノードからの アクセス)や，計算終了後に不要なデータ(上書き可)など，容量に比べて高I/O負荷になるデータに対する対処。

◆モデル案　・電力削減のため，基本はオンライン(第二階層)とオフライン(第三階層)の二階層。
・負荷の高いI/Oを受けとめるための，キャッシュとしての第一階層。
・各階層のストレージ容量(S_1～S_3)と，各層間の必要帯域(B_a～B_d)を定義。

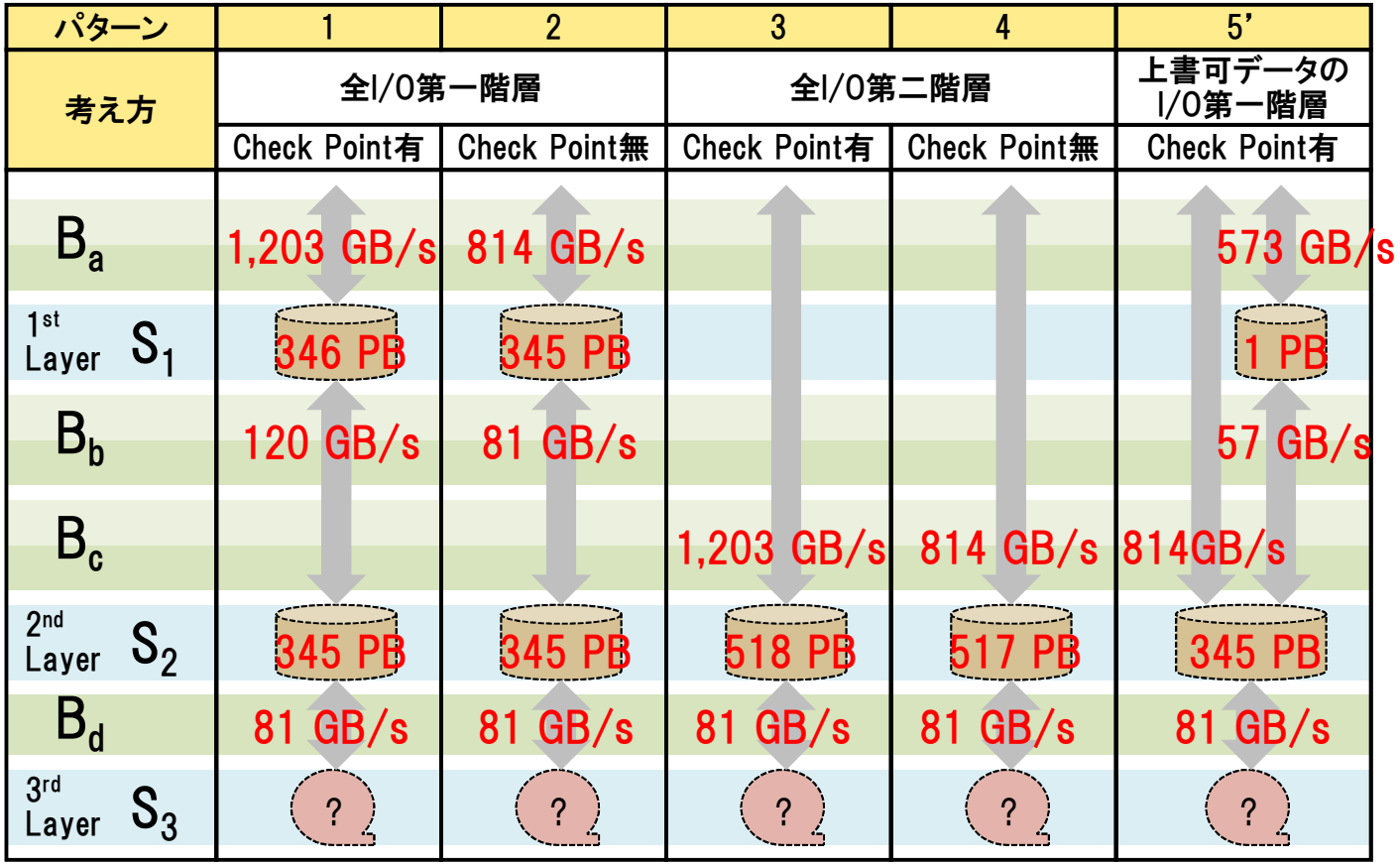


→・構成案提案に向けては，アプリI/O要件(スライド4-2)をS_1～S_3，B_a～B_dに反映

付録3-3. ご回答活用例 -I/O要件の構成モデルへのマップ-

HITACHI
Inspire the Next

■時系列シミュレーション型のアプリの要件(最低限必要な容量, スループット)



HITACHI
Inspire the Next

END

計算科学ロードマップからの計算機資源要件抽出
～特にI/O部分について～

2013/12/25

株式会社　日立製作所


HITACHI
Inspire the Next