Vol.2013-HPC-141 No.22
2013/10/1

エクサスケールスパコンに向けた耐故障性の評価 — TSUBAME2.0 を例にして —

松岡 聡[†1] 佐藤 賢斗[†1,2] 遠藤 敏夫[†1]

過去には大規模マルチペタスケールスパコンはクラスタ型では障害の頻出により絶望的で、専用設計でないと実現しないとも言われていた。しかし、今日では TSUBAME2.0 に代表されるクラスタ型のスパコンは 100%近い負荷でも問題なく実運用に供している。これは従来のサーベイは信頼性の低いクラスタの経験値によるラフな算出で、詳細なフォルトモデルが欠落していたからであり、故障発生率を過大評価していたことと、故障に対する種々の予防および耐故障性の種々のテクニックを考察していなかった事に起因する。TSUBAME3.0 からさらにエクサスケールに向け、我々はその評価を行っており、そこから何故 TSUBAME2.0 は成立するのか、将来エクサに向けてどうか、を論ずる。

Evaluating Resilience Towards Exascale --- Tsubame2.0 as an Example ---

Satoshi Matsuoka[†1] Kento Sato[†1,2] Toshio Endo[†1]

Past research has conjectured that multi-petascale supercomputers will not be feasible with clusters due to frequent faults, and only dedicated architectures would be the solution. However, circa 2013 TSUBAME2.0, a multi-petascale cluster, is operating fine even with nearly 100% load in busy seasons. This is largely due to (a) overestimation of faults due to early data derived from inadequately designed nodes, and lacked fine-grained assessment of the component faults, and (b) it did not accommodate for various fault prevention and recovery techniques. We are evaluating the feasibility of fault tolerance methodologies based on TSUBAME2.0, and will discuss the feasibility1 of the approaches towards TSUBAME3.0 in 2015 and beyond towards exascale.

1. はじめに

近年 2020 年のエクサスケールコンピューティングの機運が高まるにつれ、そのシステムの大きさおよび複雑さゆえに耐故障性に対する関心が高まりつつある。特に DARPA による Exascale Report [2]の分析によれば、10 億並列を実現するエクサスケールのマシンは 3 分に一度障害が起こるとの指摘している。実際は種々の方策によりそこまで極端に低信頼にはならないようなハード・ソフト面の対応の研究が盛んだが、そもそも大規模なスパコンがどのような故障モデルを持ち、どの程度の信頼性があるのかはあまり明らかになっていない。特に、複雑な障害は単純なハード故障の FIT 値のみでは算出できず、システム階層をまたがった、ソフトウェアの障害を含めた複雑な挙動を示す可能性がある。

特に、耐故障性のソフトウェアは、その動作パラメタを決定するために、種々の基本的な MTBF 等のパラメタが必須である[3,4,5]。例えば耐故障性のある MPI で実行する際には、どの程度の間隔でチェックポイントをとるのが最適であるか、というのは確率モデルより算出されるが、そのためには基本的な故障率のデータが必要である。しかしながら、IBM BlueGene/L において幾つか見られるものの[6]、より複雑なノード構成を要するクラスタ型、特にアクセラレータを搭載し、それぞれのノードがマルチテラフロップスに及ぶタイプの、いわゆる”Fat Node”型のスーパーコンピュータでは殆どデータがないに等しく、ある発表資料では RAS にかかる費用が全体の 1/4 以上を占めるという信じがたい数字まで根拠なく示されている[1]。

そこで、我々は TSUBAME2.0 の稼働時から蓄積されている故障ログ[7]を元に、TSUBAME の故障モデルや MTBF を全体およびコンポーネンツ単位で算出することを試みている。今のところ中途の段階だが、以下のような結果が出ている：

(1) TSUBAME2.5 において解決される TSUBAME2.0 のアーキテクチャ上の不具合を除くと GPU の信頼性は上記の指摘に反し CPU+メモリに比べてもそれほど低くなく、従って FLOPS やメモリバンド幅あたりの信頼性は数倍高い。

(2) 安定状態では実質年間 500 件程度の障害が発生しているが、半数以上はノードが停止する fail-stop ではなく、ファンなどの多重性のあるコンポーネンツか、簡単な操作で回復できるソフトウェア上のエラーであり、結果として TSUBAME2.0 全体の Fail Stop の MTBF は 1.7 日程度である。

(3) 複数のノードが影響を受ける fail stop は計測期間中わずか 3 件で、個別ノードの fail stop の 1/70 程度である。これにより、TSUBAME のような fat node なクラスタアーキテクチャでは、後述するようにローカルチェックポイントを行う階層的チェックポイント手法が有効

†1 東京工業大学
Tokyo Institute of Technology
†2 日本学術振興会特別研究員
JSPS Research Fellow

©2013 Information Processing Society of Japan

であると予測される。

(4) 障害の一つとしてノードがブートしないのが全体の 1/5 ほど見受けられる。これは TSUBAME2.0 が現在夏季昼間のピーク抑制運転によるものと推測され、今後の対策が求められる。

これらのデータおよび解析結果は TSUBAME2.0 のようなクラスタ型の大規模スパコンが[1]の主張と異なり、高信頼できちんと運用できる証拠となっている。実際、TSUBAME2.5 はほぼ同構成のマシンでピーク性能においては 5.7 ペタフロップスであるが、倍のピーク性能を持つ京コンピュータは、Google および TSUBAME における DIMM の自然故障率[8]データから、その 70 万本にものぼる DIMM だけでも全体の MTBF は 24 時間以下であると予測され、それに対し十分に伍する信頼性を獲得しているといえよう。

ただし、これらの信頼性を得るのは単純な運用では困難で、Single point of failure を極力排除する高信頼なアーキテクチャとともに、障害を事前に検知し、システムが自動対処して、実質的にユーザには見せないメカニズムが重要である。TSUBAME2.0 はそれらに関しても技術的に毎年蓄積・改善しており、全体の故障率を下げるのに役立つ。

本稿では、障害解析の現状を述べるとともに、TSUBAME3.0 の障害という観点からのフィージビリティ、およびエクサスケールへの妥当性を検証する。

2. TSUBAME2.0 スーパーコンピュータ

TSUBAME2.0 は、東京工業大学学術国際情報センターにて設計、NEC/HP/NVIDIA などの企業連合体と共同開発され、2010 年 11 月に運用開始されたスーパーコンピュータである。国内で初のペタフロップス越えを達成し、2010 年 11 月の Top500 ランキングにおいては世界 4 位、同 Green500 ランキングにおいては世界 2 位および“Greenest Production Supercomputer”賞を獲得した。図 1 にシステム構成の概要を示す。その主な特徴は下記の通りである。

高演算性能・高バンド幅を持つ計算ノード：主要部は Thin ノードと呼ばれるノード 1408 台から成り、各ノードはマルチコア CPU2 基と GPU3 基、およびメモリ 54GB または 96GB を搭載した HP 社 HP Proliant SL390s G7 である。CPU としては Intel 社製 6 コア Xeon (Westmere-EP) 2.93GHz を搭載し、GPU としては NVIDIA 社製 Tesla M2050 を搭載する。各 M2050 GPU には 448 個のコア、および 3GB のメモリが搭載され、その理論ピーク性能は 515GFlops である。Thin ノード一台の理論ピーク性能は約 1.7TFlops に達する。CPU と GPU 間のデータ通信オーバーヘッド削減のために、導入当時で最速の通信路である 8GB/s の PCI Express 2.0 x16 を採用する。

上記の Thin ノードに加え、ノード内で特に大容量の共有メモリを必要とするアプリケーションのために、Medium/Fat ノードと呼ばれる HP 社 Proliant DL580 G7 ノードを 34 ノード持つ。ノードあたりのメモリ容量は、128GB、256GB または 512GB である。

階層型大規模ストレージ：TSUBAME2.0 のストレージは、様々な I/O 要求を行う多種アプリケーションのため階層化され、(1) ノードローカル SSD、(2) ディスクによる共有ストレージ、(3) テープライブラリから成る。

(1) 各計算ノードはローカルストレージとして、ハードディスクの代わりに 120-240GB の容量の solid state drive (SSD)を持つ。チェックポイントなどの用途でこの SSD を利用することにより、共有ストレージへの負荷を軽減する。

(2) 共有ストレージは合計 7.2 ペタバイト(PB)の容量(raw capacity)を持つ。このストレージは 6 つのファイルシステムに分割され、1 つがホーム領域、5 つが並列ファイルシステム領域として利用される。各ファイルシステムにおいてデータを蓄積するのは DDN 社 SFA 10000 ストレージシステムである。ホーム領域は高信頼性に焦点をおきつつ、高速な NFS 性能、および CIFS, iSCSI などの複数プロトコルに対応する。並列ファイルシステム領域はスケーラビリティを焦点に設計されており、ファイルシステムとして Lustre および GPFS を採用する。

(3) さらにバックアップを主目的とした三次領域として、計 8PB(非圧縮)の既存の SL8500 テープライブラリと接続されている。階層的ファイルシステムの利用により、このテープライブラリと並列ファイルシステム領域間で透過的なデータアクセスを提供する。

これらストレージシステムは、過去の TSUBAME1.0 においてストレージがしばしば Single Point of Failure になった経験から、大幅に多重化され、かつ自動回復されるようになっている。例えば、ストレージに対する Infiniband でのアクセスパス・コントローラ・サーバ群は全て多重化されており、かつ障害時には自動的に failover する。HDD は RAID6 であることは勿論で、かつ個体 HDD の障害時には自動的にスペアドライブからリビルドが行われる。ファイルシステムは NFS, Lustre, GPFS と多重化され、致命的なバグ時にもシステムソフトウェアレベルで多重化されている。後の障害記録が示すように、ストレージ関係の障害は SSD を除くと一年間でわずか 17 件であり、かつ Fail-stop は一件もなく、高信頼を具現化している。

フルバイセクションネットワーク：1400 以上の計算ノードおよびストレージを結合する高速インターコネクトとし

©2013 Information Processing Society of Japan

て、4x QDR InfiniBand を採用している。インターコネクトはほぼ同形の二つの InfiniBand ネットワークから成り、それぞれがフルバイセクション・ファットツリーと呼ばれるトポロジを成す。ファットツリー構造の上部にはコアスイッチとして 324 ポートの大規模 InfiniBand スイッチを計 12 個採用し、コアスイッチと下部のエッジスイッチ間は光ファイバーで結線されている。ファイバーの本数はシステム全体で 3500 本、総延長 100km 程度である。

現行の Infiniband は adaptive routing はサポートされていないが、上位の光リンクの障害発生時にはルーティングテーブルが subnet manager によって自動的に障害を回避するように書き換えられる。ただし、折角 TSUBAME2.0 は二つの IB リンクをノード毎に持ち合わせていて、別々のエッジスイッチに接続されているにも関わらず、上位レイヤの MPI は OpenMPI, MVAPICH など全て両リンクの利用が前提で、シングルリンクへのフェイルオーバーがない。これは課題ではあるものの、実際の Infiniband の完全 fail-stop 障害は一年間で高々数回であった。

2013 年 9 月に TSUBAME2.0 は、最新の Kepler 世代の Tesla K20X へ GPU アクセラレータを置き換えることにより、理論ピーク性能 5.76PFlops(倍精度)、17PFlops(単精度)を持つ TSUBAME2.5 にアップグレードされた。本稿では、TSUBAME2.0 時代の障害履歴に基づく評価を行うが、この結果はそのまま TSUBAME2.5 にも当てはまると強く予想され、TSUBAME3.0 および将来のエクサスケールマシンへの設計時の参考値となる。

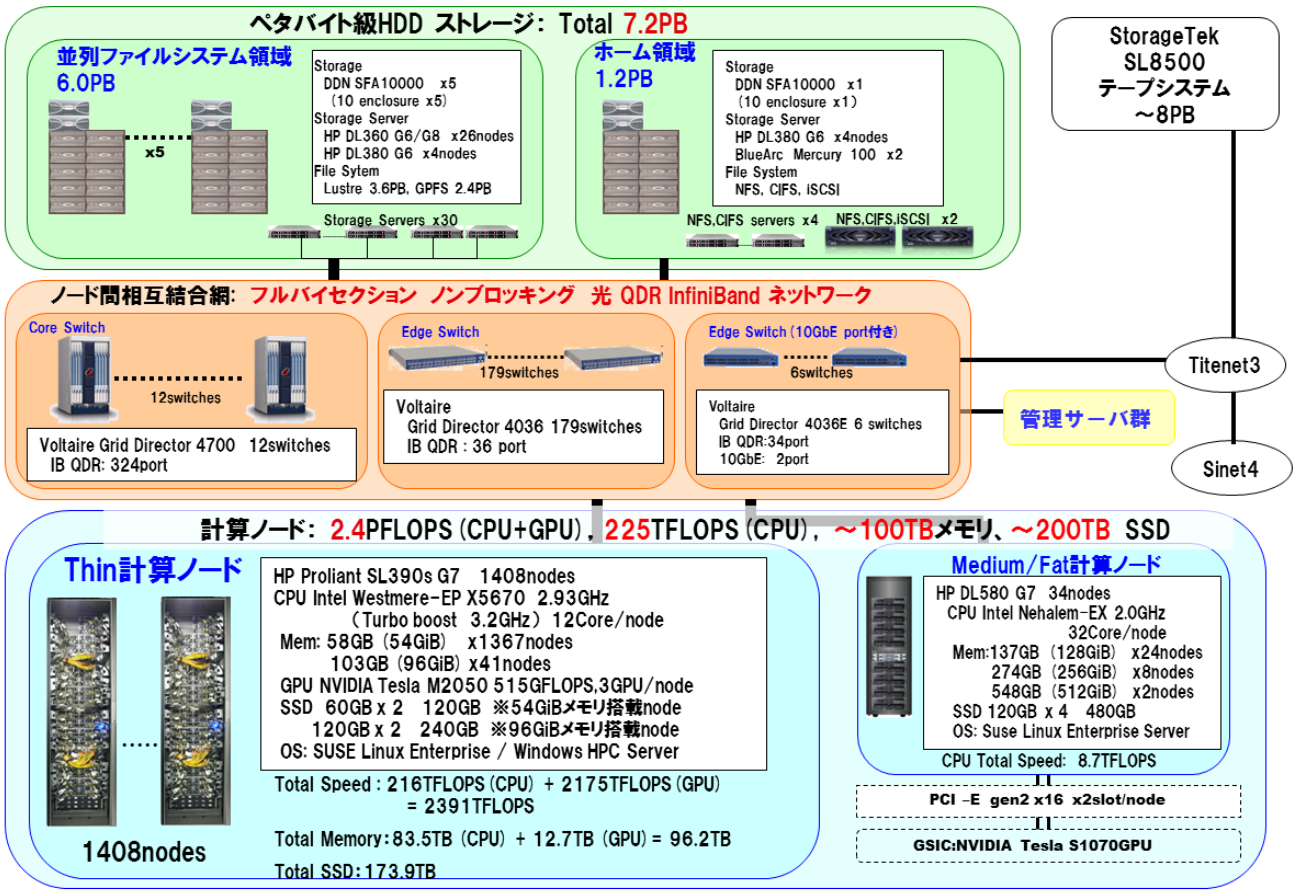

図 1 TSUBAME2.0 システム構成図
Figure 1 System structure of TSUBAME2.0

3. 解析手法

TSUBAME2.0 は通常の RAS も記録されるが、それに加えて障害の日時や場所、その種類や対処法などを記したデータベースを稼働時から Web 上に公開している[7]。記録される項目は、ホスト(ノード名)/機器、キュー、発生日付、復旧日付、障害状況、原因、対処、影響範囲、(障害の)カテゴリ、である。TSUBAME の調達契約において、このデータベースを遅滞なく維持管理することが求められており、障害の発見や対処を行う度に担当エンジニアが手動で更新している。後述のように自動チェックも頻繁に行われているが、これらの結果も最終的には手動で入力され、自動化はされていない。生の RAS データを人間可読な方式に自動変換していくのは今後の課題である。

©2013 Information Processing Society of Japan

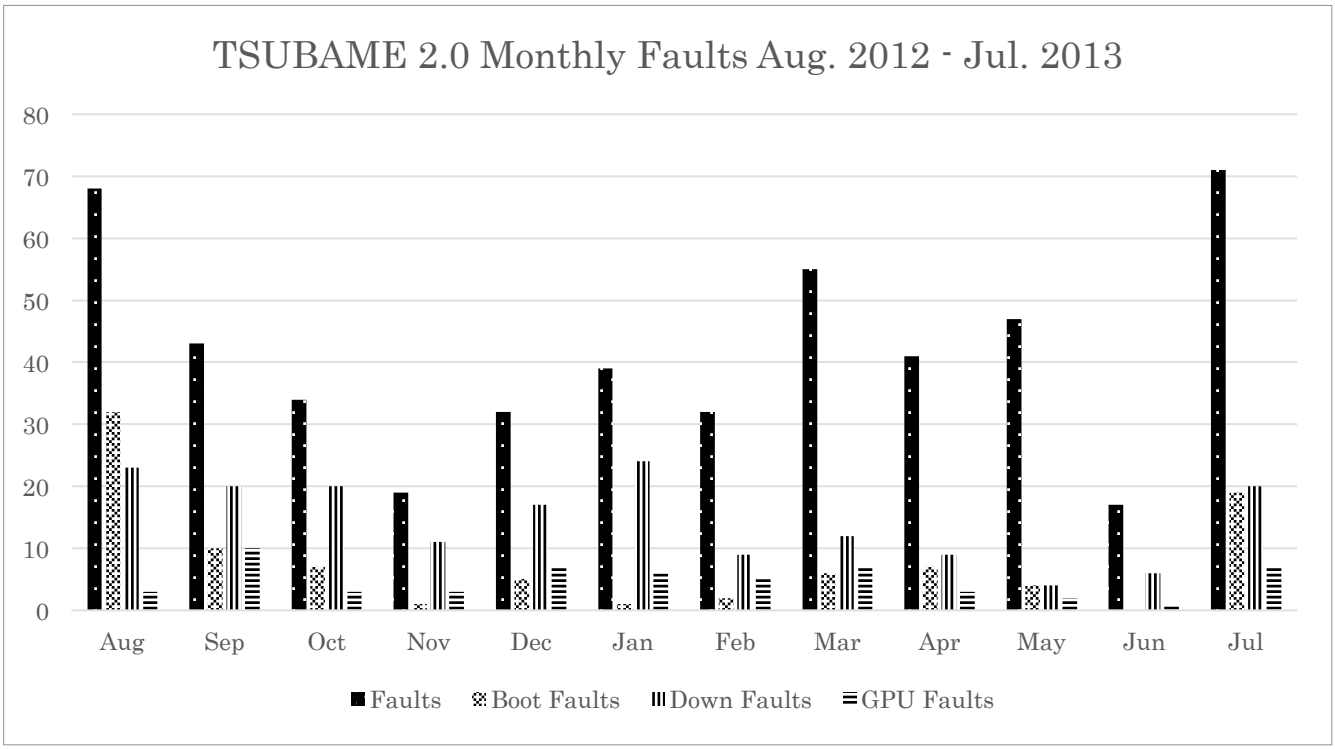

図 2 月別障害報告

Figure 2 Monthly Faults Statistics

2013 年 8 月現在、TSUBAME の障害データベースのエントリは 2 年 10 か月の運用で 3000 件近い。しかしながら、検討の結果、このデータはありのままでは使えず、相当なキュレーションが必要な事が判明した。以下が主な理由であるが、今後の障害ログの記載法にも大いに参考になった。

(1) 障害エントリのミス、一貫性の欠落: 種々な理由で、データベースのエントリの一貫性が欠落し、同じ障害や対処法でもエントリが異なることがある。例えば、原因が単純に「ノードダウン」とそのままなっている場合と、CPU 障害と判明した場合はその旨の記載がある。また、明らかに SSD の障害なのに障害項目は GPU となっているなどの記載ミスも数々見られた。これらは一通り見渡した上で類型化し、ほぼ同一とみられる現象を合算する作業を要した

(2) システムの経年変化による障害の質的変化: 良く言われるのは、システムは初期不良が多発する初期、安定する中期、コンポーネンツの老朽化により再び故障率が上昇する末期と変化する事だが、それだけではなく、システム運用や以前は不明だった障害への有効な対処法、更にはシステムの使用法の変化など、種々の要因が挙げられる。そこで、今回は初期不良を克服し、安定稼働時の一年案である 2012 年 8 月 1 日から 2013 年 7 月 31 日までの一年間のデータを用いることとした。これは 2013 年 8 月より TSUBAME2.5 への入れ替えが開始され、安定したデータが取得できるギリギリのところをデータとした事情がある。

(3) 情報の欠落：先に述べたように、障害の結果として fail-stop 障害が起きたのか、あるいは問題はなくジョブは継続したのか、あるいは障害はブート時だったのか、などの情報が今回の解析には重要だが、それらの情報が欠落しており、他の障害情報から補完する必要がある項目が多々あった。

(4) 加えて TSUBAME、特に GPU に関する固有の事情があった。表 1 にある通り GPU-PCI の項目には 362 件の Fail-stop 障害が報告されている。しかしながら、これらに対するデータベースに記載されている対処はほとんどが GPU の位置のスワップや抜き差しであり、部品交換を伴っておらず、最終結果からは除外してある。これは以下の理由による：抜き差し等で復活するのは一見接触不良に見受けられるが、実は全く違う理由であることが 2013 年の TSUBAME2.5 へのアップグレードの設計およびテスト時に判明した : TSUBAME2.0 に装着された初期ロットの M2050 カードは、瞬間的には定格の 225W を大きく上回る電力を消費するため、TSUBAME の計算ノードの e-fuse という保護回路がトリップしてしまうことが判明した。E-fuse トリップは Uncorrectable PCI Express Error として検出され、ハードウェアエラーとしてサーバー管理システム（iLO）に記録されて OS はハングし、電源を落とすコールドリセットで復活する。実際、GPU-PCI と記載されている全ての項目はこれである。今回の TSUBAME2.5 への

©2013 Information Processing Society of Japan

表 1 コンポーネント毎の障害解析
Table 1 Failure analysis of components

	SUM	Boot Failures	Fail-stop Failures	Unaffected	Multi Failures	Repairs	GPU Repairs	DIMM Repairs
Unknown Boot Failure	35	35				0		
CPU	16	4	12			14		
Disk/Storage (wo SSD)	17	5		12		5		
Unknown Node Failure	15		15			0		
Fan	100			100		13		
GPUs								
GPU-PCI	362		362			96	39	
GPU-Link	16			16		10	3	
GPU-ECC	10		10			10	10	
GPU-Unknown	10	6	4			9	5	
Memory	12	1	7	4		11		14
Network								
Infiniband	22		4	16	2	4		
Other Networks	78		55	23		0		
Other HW	27	22		5		24		
Batch System (PBS Pro)	13		3	10		0		
PSU	33		26	7		33		
Rack	6			5	1	4		
SSD	34	12	22			32		
System Board	22	9	13			22		
Total Corrected Total	828 498	94 94	533 210	198 191	3 3	287 209	57 57	14 14

(5) GPU 入れ替えにともない、e-fuse 搭載部品もマージンを改善したものに入れ替え、全てのストレステストでも再発していない。よって、TSUBAME2.5 を含む他のマシンでは本来発生しない、TSUBAME2.0 特殊事情と判断し、表の Corrected Total では GPU を実際に交換した場合以外は GPU-PCI と GPU-Link の障害数は除外して正規化してある。

(6)

これらのキュレーションを施したデータは別途公開する予定である。問題は、不完全情報や作業ミスで若干であるがキュレーション時にエラーが入り込む余地があることである。総論では、障害の項目数が多いので全体の結果には影響は少ないといえよう。しかしながら、例えば上記の GPU でも、2012 年秋ごろまではスワップなどでなくより積極的に予防交換していて、GPU 障害を多く見積もりすぎている可能性がある。今後も TSUBAME2.5 での新しいデータを用い、精度を改善していく所存である。

4. 解析結果

現状では解析の中途であるが、現状まで判明している結果を以下に列挙する：

(1) 3 節のキュレーションの結果からは、TSUBAME2.0 の信頼性は現状の安定状態では比較的高い。実質年間 500 件程度の障害が発生しているが、Fail-stop は半数以下の 210 回であり、半数以上はファンなどの多重性のあるコンポーネンツか、簡単な操作で回復できるソフトウェア上のエラーである。結果として TSUBAME2.0 全体の Fail Stop の MTBF は 1.7 日程度であり、1432 ノードの各ノード単位の MTBF は約 2500 日間であり、5.7 ペタフロップスの TSUBAME2.5 においてもそれはそのまま継承される可能性は高い。後述するが、京コンピュータの全体の MTBF はメモリだけでも 24 時間を切るので、十分に信頼性は高いと言えよう。

(2) TSUBAME2.5 において解決される TSUBAME2.0 のアーキテクチャ上の不具合を除くと GPU の信頼性は上記の指摘に反し CPU+メモリに比べてもそれほど低くなく、従って FLOPS やメモリバンド幅あたりの信頼性は数倍高い。CPU+メモリの fail stop 障害数は年間 19 件、予防措置を含めた交換は 25 件であり、一方 GPU はそれぞれ 53、57 件である。これは先に述べたように 2012 年秋の段階で GPU-PCI のエラーを予防的に交換していた時期を含んでおり、GPU の障害数は大目に換算されていると言え、図 2 の月別の GPU 障害数でも、2013 年度になり明らかに GPU 障害が減少しているのが判明している。仮にこのままとしても、TSUBAME2.0 は 2952 個(ソケット)の Intel Westmere Xeon CPU、4264

©2013 Information Processing Society of Japan

個の NVIDIA Kepler M2050(一部 M2070)GPU があるので MTBF はおおむね 118 年対 75 年とそれほど変わらない。一方ピーク演算数であたりのエラー率は 1.61^{19} FLOP 対 2.22^{18}FLOP と、7 倍以上の開きがあり、GPU が圧倒的に有利である。TSUBAME2.0/2.5 の CPU＋GPU の合算数は 7216 個なので、TSUBAME3.0 が同数かそれより少し多い程度であれば、現状と同等の MTTI や MTBF が十分確保できると考えられ、20-30 ペタフロップスの性能でも十分な運用が可能である。

(3) 複数のノードが影響を受ける fail stop は計測期間中わずか3件で､個別ノードの fail stop の 1/70 程度である。また、表 1 から、Infiniband や並列ストレージの障害は非常に少なく、SSH なども今後の障害検出だけでなく復旧や Ether も併用したロバスト化によって十分対処できる範囲内であると結論づけられる。

これにより、以下が結論付けられる：TSUBAME のような fat node で、かつ電源・ネットワーク・ストレージなどが多重化などでかなりロバストになっているクラスタアーキテクチャでは、BlueGene のような thin node､専用ネットワークで RAS ソフトウェア層が個別ノード単位では薄いのと異なり、障害は個別ノードで閉じており、複数ノードで障害が連鎖的に伝わることは(その MPI ジョブが停止する以外には)ほとんどない。よって、ほとんどの場合はローカルチェックポイントを行う階層的チェックポイント手法が有効である。

(4) メモリの交換は僅か 14 モジュールであり、かつ GPU の ECC エラーは 10 件である。TSUBAME の DIMM の総数は約 17,000 本であり、DUE(Double-bit Uncorrectable Error)は 0.082%であって、[8]における Google の調査の結果である 0.22%のわずか 1/3 近い。これは、安定期であること、スパコンなのでクラウドと比較してより高品質な DIMM を使っていること､および冷却が高効率で温度湿度環境が良い事などが複合的に要因となっていると予想される｡一方 GPU のメモリエラーはボード毎に年率 0.23％だが、これは[10]らの論文による BG/P の 40,960 ノードある各ボードのメモリの Chipkill エラーを年換算した 0.83%より大幅に低い。これからも、[1]の GPU クラスタに対する指摘が TSUBAME2.0 のような専用設計のスパコンには全く当てはまらないどころか、ずっと高信頼であることがわかる。

一方､京コンピュータは TSUBAME2.0 と同等の DIMM の DUE だと仮定すると、DIMM の総本数は約 70 万と TSUBAME2.0 の約 40 倍あるので、年間の故障本数は約 600 本近い。つまりメモリの DUE だけでも京の MTBF は半日もないと予測される。京の Linpack は 26 時間完走したが、それは全メモリを使っていないのが一つの大きな理由であり、実運用でも同様であるが、全系で MTTI が一日を超えることは考えにくい。このあたり、TSUBAME と異なり現在京の障害報告は公開されていないので、詳細は不明である。

(5) 図 2 における毎月の障害数を鑑みると、7-9 月に障害が増えている。これは冷却が問題なのでなく、以下の理由が障害の一つとして考えられる。ノードがブートしないのが全体の 1/5 ほど見受けられるが、TSUBAME2.0 は現在夏季昼間のピーク抑制運転を毎夏 7-9 月ごろ実施しており､9:00-20:00 の間 1/3 程度のノードを毎日完全自動停止する。これらが夜になっても起動しない障害が全体のブートエラーの 1/2 程度を占めている。既にタイミングエラーなどによるソフト的エラーはリトライなどで克服しているので、これらはそれにも拘わらずブートしない障害であり、今後の対策が求められる。

これらのデータ9から､TSUBAME3.0 は TSUBAME2.5 から踏襲する後述の事前障害予防措置などのみでも、京や BlueGene/Q を遥かに上回る十分高信頼なシステムとなると予想される。更に我々が研究している種々のソフトウェアレベルの耐故障性のライブラリ[9,**Error! Bookmark not defined.**],より、(同じパラメタで)速度をほとんど犠牲にせずに遥かに高信頼となることがわかる[5,11]。ただし、カギとなる SSD は、TSUBAME3.0 では大幅にユニット数が増加するだけでなく、テクノロジ・利用形態・Wear Leveling のアルゴリズムなど、非常に多くの項目が変わるので、別途評価が必要である。

一方エクサスケールでは事情が異なる。TSUBAME4.0 が同様の構成で達成できるのは 2020 年においても 100-200 ペタフロップスであり､エクサスケールには 5-10 倍程システムサイズが増加する必要がある。すると､現在 40 時間程度のシステム MTBF が(流石に数分ではないものの)4-8 時間程度に短縮されてしまい、上記の耐故障ソフトウェアを含む対処が重要になる。別な見方からすれば、全系を用いるアプリケーションから見える MTBF は､TSUBAME4.0 において 200-400 時間にならないとエクサで同等の運用はできないことになるが、それは現行のハードウェアの高信頼化だけでは限界10があり、ソフトウェアによる対処が必須となる。

©2013 Information Processing Society of Japan

表 2 TSUBAME2.0 障害診断リスト
Figure 2 TSUBAME2.0 Health Check List

Check Category	Check Performed	Interval	Action on Fautl	Subject	Av. Exec Time
Network	Infiniband Status, Check	2H	Notify Sysadmin	Node	5.6E-02
Clock	System Clock Drift	2H	Notify Sysadmin	Node	2.4E-01
GPU	PCIe Link Speed, Driver Permission, Device Memory ECC Error	2H	Auto Offline	Node	7.8E-02
HDD	Available Space, Filesystem Mount	2H	Notify Sysadmin	Node	1.2E-02
SSD	Partition and Size	2H	Notify Sysadmin	Node	Ditto
SSD	Permission	1D		Node	Ditto
SSD	fsck /scratch space	1H	Notify Sysadmin	Node	Ditto
SSH	SSH login deamon	1H	Auto Offline	All Nodes	2.3E+01
Process	Zombie Process	1H	Kill Zombie	Node	6.2E+00
PBS	PBS scheduler status, qstat response (60 seconds)	1H	Notify Sysadmin	Admin Node	2.3E-01
PBS	MOM Check			Node	5.4E-02
PBS	Decommision Waiting Reserve Job	1H	Auto Decommisioning	Admin Node	6.5E+00
OpenSM	Check operation	1H	Notify Sysadmin	Admin Node	7.7E+01
Lustre	Check MDS, OSS, OST activity	1H	Notify Sysadmin	Admin Node	1.5E-01
Interactive	Load Average	1D	Notify Sysadmin	Interactive	8.0E-03
H (Reservation) Queue	Check Actual reservation and batch status	1D	Notify Sysadmin	Admin Node	4.4E-01
VM Check	SSH Login, available space, etc.	1D	Notify Sysadmin	All Virtual Nodes	3.0E+02
IBCORE/IBEDGE	Link up/down, link speed	1D	Notify Sysadmin	Admin Node	2.5E+01
IBEDGE	connectivity to storage	1H	Notify Sysadmin	Admin Node	8.8E-01

5. TSUBAME2.0 における事前障害予防措置

先に述べたように、TSUBAME では単にユーザのアプリケーションが障害に遭遇した時や、ノードが fail-stop で落ちた時だけでなく、常時監視プログラムがシステムの様々な側面をチェックしていて、一部の障害には自動対処している。表 2 に現在の一覧を示すが、頻度、種類、システムのどの部分をチェックするか、など、バラエティに富んでいる。これらのチェックはユーザのアプリケーションが実行されていても強制的に実行され、ユーザは止めることはできない。一方実行時間は頻度に対して十分短く、他の OS デーモンなどと比較してもユーザの実行には影響を与えないはずだが、今後システムの大規模化に伴い性能劣化のノイズとならないように注意する必要がある。

これらのチェックで検出されるエラー数は多く、例えば Zombie プロセスの消去や SSH, VM の状況チェックと修正機能で、システムの安定化に大きく寄与している。

6. おわりに

我々は TSUBAME2.0 の稼働時から蓄積されている故障ログを元に、TSUBAME の故障モデルや MTBF を全体およびコンポーネンツ単位で算出することを試みた。これらのデータおよび解析結果から TSUBAME2.0 のようなクラスタ型の大規模スパコンが種々の主張と異なり、高信頼かつ安定した運用ができていることを示した。

今後は、BlueGene のような Low レベルの RAS の自動収集記録機構と、現行の TSUBAME の Middle レベル RAS チェック、さらに SE などの人間を介在させた、より柔軟かつ High レベルな RAS チェックを統合し、複合型障害記録機構を実現する。また、これらの故障率のデータとシステムアーキテクチャからより詳細な故障モデルを構築する。さらに、それを11使ってハード・ソフトでの対策が TSUBAME3.0、更には TSUBAME4.0 やエクサスケールマシンで有効か検証するとともに、我々が研究してきた耐障害 API をエクサスケールに対応させ、高信頼アプリケーション実行環境を実現する。

謝辞　本研究は日本学術振興会 基盤研究 S 23220003 の助成を受けたものです。

参考文献

1) S. Chari : “IBM blue Gene/Q: The most energy efficient green solution for high performance computing,” http://public.dhe.ibm.com/common/ssi/ecm/en/dcl12350usen/DCL12350USEN.PDF, Jun. (2011)
2) K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. S. Williams, K. Yelick, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Keckler, D. Klein, P. Kogge, R. S. Williams, and K. Yelick :“ExaScale Computing Study: Technology Challenges in Achieving Exascale

©2013 Information Processing Society of Japan

Systems Peter Kogge, Editor & Study lead," (2008).
3) N. H. Vaidya : "On Checkpoint Latency," College Station, TX, USA, Tech. Rep., (1995).
4) Jitsumoto, H. Endo, T. Matsuoka, S. : "Environmental-aware optimization of MPI checkpointing intervals", 2008 IEEE International Conference on Cluster Computing, pp.326-329, Sept. 29 2008-Oct. (2008).
5) Kento Sato, Naoya Maruyama, Kathryn Mohror, Adam Moody, Todd Gamblin, Bronis R. de Supinski, and Satoshi Matsuoka. : Design and modeling of a non-blocking checkpointing system. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '12). IEEE Computer Society Press, Los Alamitos, CA, USA, , Article 19 , 10 pages. (2012)
6) Liang, Y. Zhang, Y. Jette, M. Anand Sivasubramaniam and Sahoo, R. : "BlueGene/L Failure Analysis and Prediction Models," International Conference on Dependable Systems and Networks, 2006. DSN 2006., pp.425,434, 25-28 June (2006)
7) "Tsubame 2.0 - monitoring portal," http://mon.g.gsic.titech.ac.jp/.
8) B. Schroeder, E. Pinheiro, and W.-D. Weber : "DRAM errors in the wild: A Large-Scale field study," in SIGMETRICS (2009)
9) Bautista-Gomez, L., Komatitsch, D., Maruyama, N., Tsuboi, S., Cappello, F. and Matsuoka, S.: FTI: high performance Fault Tolerance Interface for hybrid systems, Proceedings of the 2011 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, Seattle, WS, USA (2011).
10) A. A. Hwang, I. A. Stefanovici, and B. Schroeder, "Cosmic rays don't strike twice: understanding the nature of DRAM errors and the implications for system design," in Proceedings of the seventeenth international conference on Architectural Support for Programming Languages and Operating Systems, ser. ASPLOS XVII. New York, NY, USA: ACM, 2012, pp. 111–122. (2012).
11) Moody, A., Bronevetsky, G., Mohror, K. and de Supinski, B. R.: Design, Modeling, and Evaluation of a Scalable Multi-level Checkpointing System, Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, SC '10, Washington, DC, USA, IEEE Computer Society, pp. 1–11 (online), DOI: 10.1109/sc.2010.18 (2010).
12) Gomez, L. B., Ropars, T., Maruyama, N., Cappello, F. and Matsuoka, S.: Hierarchical Clustering Strategies for Fault Tolerance in Large Scale HPC Systems, Proceedings of the 2012 IEEE International Conference on Cluster Computing, CLUSTER '12, Washing- ton, DC, USA, IEEE Computer Society, pp. 355–363 (online), DOI: 10.1109/CLUSTER.2012.71 (2012).

©2013 Information Processing Society of Japan