# ABCIデータセンターにおけるハードウェア障害の傾向

高野 了成[1,a)]　滝澤 真一朗[1]　三浦 信一[1,2]　谷村 勇輔[1]　小川 宏高[1]

## 1. はじめに

大規模システムの安定運用には，障害の解析とその対応が必須である．従来から大規模高性能計算機システムについては様々な障害解析がなされており，高信頼化技術の開発や運用の改善に役立てられてきた．例えば，大規模 GPU クラスタである米国 ORNL の Titan における GPU 障害に関する網羅的な解析 [1] や，京コンピュータにおける冷却がハードウェア故障に及ぼす影響の解析 [2] などが知られている．しかし，産総研が 2018 年 8 月から運用を開始した AI 橋渡しクラウド（ABCI）のように，冷却機構として高温冷却水を利用したフリークーリングシステムを採用する大規模 GPU システムは，世界的にも例がない先進的なものであり，十分な障害解析はなされていない．本ポスタでは，2018 年 7 月から 2019 年 11 月の 17 ヶ月間における，ABCI の時間的及び空間的なハードウェア障害の傾向について報告する．また，ABCI と比較的構成が近い TSUBAME3.0 [3] と比較しても遜色のないハードウェア故障率であり，その設計が妥当であることを示す．

## 2. ABCI データセンターの概要

ABCI は産総研が東京大学柏 II キャンパスに整備した，わが国の人工知能技術開発を加速するオープンで世界トップクラスの高速計算基盤であり，理論ピーク性能は 0.55 EFLOPS (FP16)，37 PFLOPS (FP64) を誇る．HPL 実効性能は 19.88 PFLOPS，14.423 GFLOPS/W を達成し，2019 年 11 月現在，TOP500 において世界 8 位および国内 1 位，Green500 において世界 6 位にランキングされている．

我々は ABCI の構築にあたって，計算機システムを収容するための AI データセンター棟（ABCI データセンター）も新たに設計・構築した．その詳細は研究会報告 [4] を参照されたい．ABCI が採用している高密度 GPU サーバ（富士通製 PRIMERGY CX2570 M4）は，1U サイズあたり 2 CPU，4 GPU を搭載する構成であり，その定格電力は 2kW 程度となる．当該サーバが 1 ラックあたり 34 台収容されており，1 ラックあたりの消費電力は 70kW に達する．ABCI データセンターは，このような高密度のサーバ機器に耐えうる電力容量，冷却能力および床耐荷重を備えている．冷却に関しては，主要な熱源であるプロセッサおよびメモリに対してウォータブロック方式による直接水冷を行い，他のコンポーネントを間接水冷とすることで，高い冷却性能とメンテナンス性を実現している．また，夏期を含めた年間を通して冷却塔のみを用いたフリークーリングシステムを用いて冷却水を生成することで，冷却に必要な消費電力の削減も実現している．この結果，年間平均で PUE 1.1 程度での運用を実現している．

[1] 国立研究開発法人 産業技術総合研究所
[2] 国立大学法人 東京工業大学
[a)] takano-ryousei@aist.go.jp

## 3. ハードウェア障害の傾向

サーバコンポーネント（CPU，GPU，メモリ，マザーボード，（GPU 接続用の）ベースボード，フラッシュディスク，InfiniBand HCA，電源ユニット等）毎に期間中の月次障害率，および障害とラック配置の関係について調査した．解析により得られた事項を下記に列挙する．

- 夏期（6/5〜9/6．高温循環水 32 °C，ラック吸気温度 35 °C 設定）の障害が多かった．サンプル数が少ないため断言はできないが，障害頻度と水温・室温には相関がある可能性が高い．
- 50%の障害は前半 1/3 のラックで起きており，故障とラック配置には弱い相関関係がある．これはジョブスケジューラ（Univa GridEngine）が先頭のラックから順にジョブを割り当てるため，利用率が高いサーバで障害が発生している可能性が高い．
- TSUBAME3.0 と比較してハードウェア交換率は変わらない．ハードウェア構成や負荷が異なるため，正確な比較ではないが，概ね同規模のシステムと比較して高い故障率ではなく，コモディティ製品を組み合わせてスパコン並みの信頼性を達成できることを実証した．

**謝辞**　本研究の一部は科研費 19H04121 の助成を受けた．

### 参考文献

[1] Tiwari, D., Gupta, S., Gallarno, G., Rogers, J. and Maxwell, D.: Reliability Lessons Learned From GPU Experience With The Titan Supercomputer at Oak Ridge Leadership Computing Facility, *ACM/IEEE SC2015* (2015).
[2] Shoji, F., Matsui, S., Okamoto, M., Sueyasu, F., Tsukamoto, T., Uno, A. and Yamamoto, K.: Long term failure analysis of 10 petascale supercomputer, *ISC2015 HPC in Asia Poster Award* (2015).
[3] 東京工業大学: TSUBAME3.0 障害履歴, `http://pm1.t3.gsic.titech.ac.jp/sysrepo/index.php` (2019).
[4] 高野了成，三浦信一，杉田　正，小川宏高，松岡　聡: 0.55 AI-EFLOPS の計算インフラストラクチャを支える超グリーン AI データセンタ，情報処理学会研究会報告 (2018).