



HPCシステムを絞りつくす

2024年12月5-6日

第24回PCクラスタシンポジウム 2024「サステイナブルなHPCに向けて」

Pacific Teck Japan 合同会社 Senior Engineer 森本 賢治

HPC/AIの社会的立ち位置

HPC・AIが単体でリソース・エネルギー収支をゼロにすることはかなり困難。

「サステイナブルな社会」を構成するための必須コンポーネントとして認知されることが必要。

多大なるリソース、予算、労力をつぎ込む以上、それを上回るベネフィットが求められる。

適切なシステムを作り、それを使い倒す、性能を絞り尽くすためのツールや手法が必要。

一方で、「サステイナブルでないHPC」は容易に想像ができてしまう。



適切なエネルギーやリソースで
相応の成果を得よう。

過剰なリソース投入や浪費を
戒めるためのフレーズ

阻害要因は何か？

ダウンタイム

- ・ システムの導入立ち上げ・メンテナンスは早急に。すぐに運用に入りたい。
- ・ 故障はモニタリングにより即時検知して対応したい。

チューニング

- ・ ハードウェア性能を完全に引き出せるアプリケーションが作れているか。
- ・ 試行錯誤にリソースを過剰に使っていないか。

待ち時間

- ・ I/O待ち、アプリケーション待ち、ネットワーク待ちがないか。
- ・ ジョブの設計は遅い

カテゴリー別 取扱製品

■ ジョブ管理システム



■ HPC向けコンテナ



■ セキュリティ分析プラットフォーム



■ ストレージソフトウェア

■ 並列ファイルシステム



BeeGFS®

BeeOND®

VAST

■ S3 オブジェクトストレージ



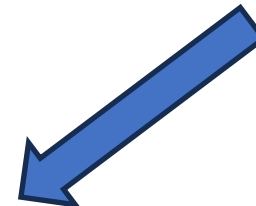
■ S3クラウドストレージサービス



■ データマネジメント



■ クラスターマネジメントシステム

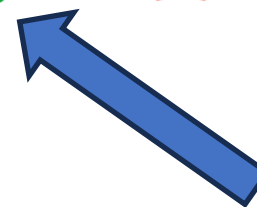


■ I/O Profiling 開発者用ツール



linaro

forge

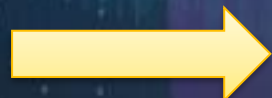
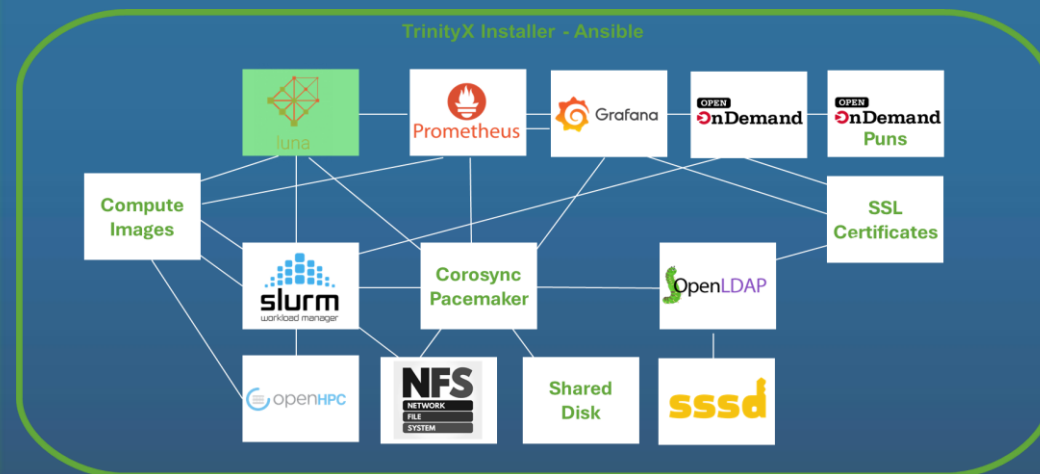
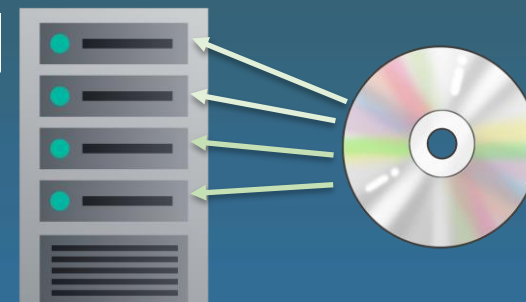


HPC/AI クラスタを構築するには

サーバーを購入するだけではクラスターは作れない。
システム構築・運用には膨大な工数がかかっています。

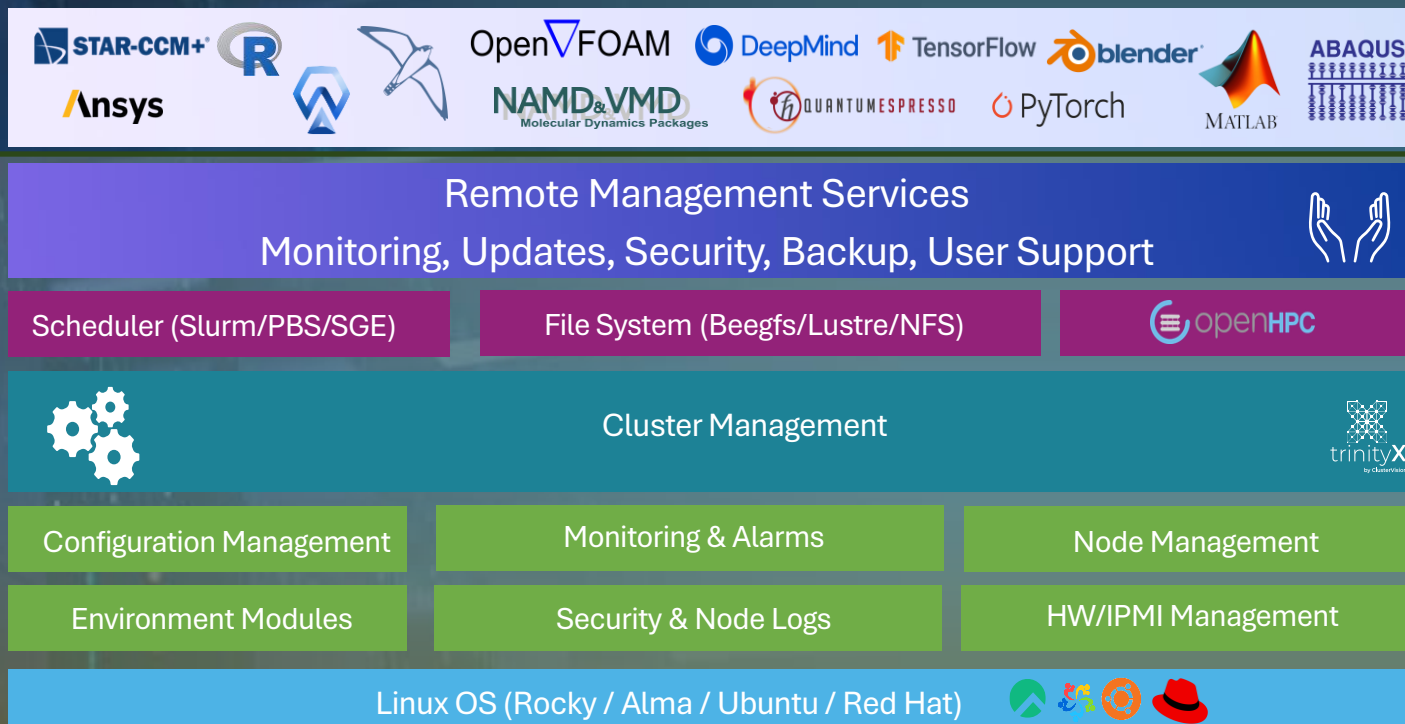
1. 全台にOSインストール、設定
2. GPU・インターコネクタ用ソフトウェア導入
3. 個別サーバーのネットワーク設定
4. 共有ストレージ作成、マウント設定
5. 共有認証基盤(LDAP等)作成、接続設定
6. ジョブスケジューラの導入、キュー設計
7. 並列計算ソフトウェアのインストール、設定
8. アプリケーションのインストール、設定
9. GUI のインストール、設定
10. 監視・モニタリングのインストール、設定
11. ユーザー管理、教育、トラブルシュートなど
12. システムアップデート、セキュリティ対応など

AMD
ROCm



コンポーネントはOSSで揃うが、システムに組み上げることは膨大な工数とノウハウ、経験が必要。
これらを自動化するツールが必須。

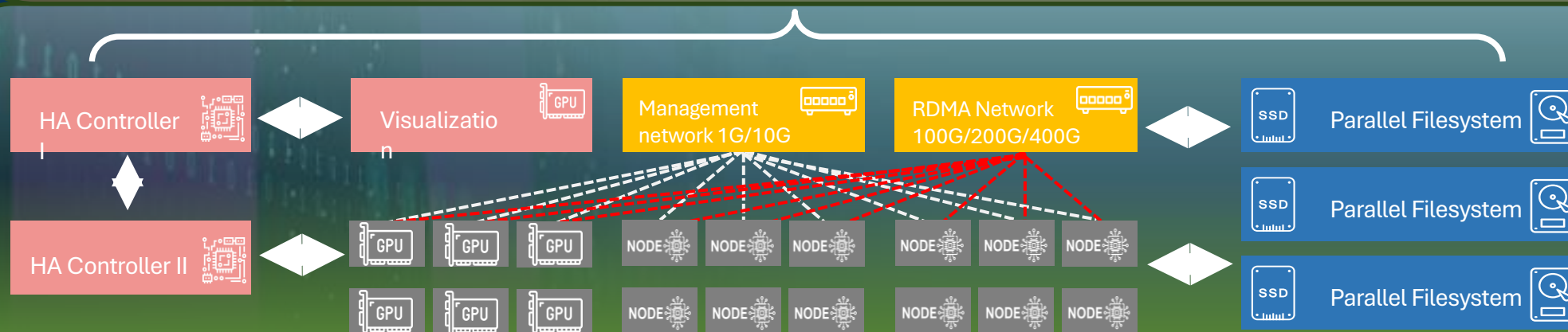
HPC・AI クラスタ(スパコン)の構造



ここを整備してハードウェアを
統合し、システムを構築・運用

様々なコンポーネント群が
個々のシステムを特徴づけてきた

⇒ 統合管理ツールを使いたい



ハードウェア群
管理ノード
計算ノード
ストレージ
ネットワーク
.....

柔軟なシステム構築・マネジメント

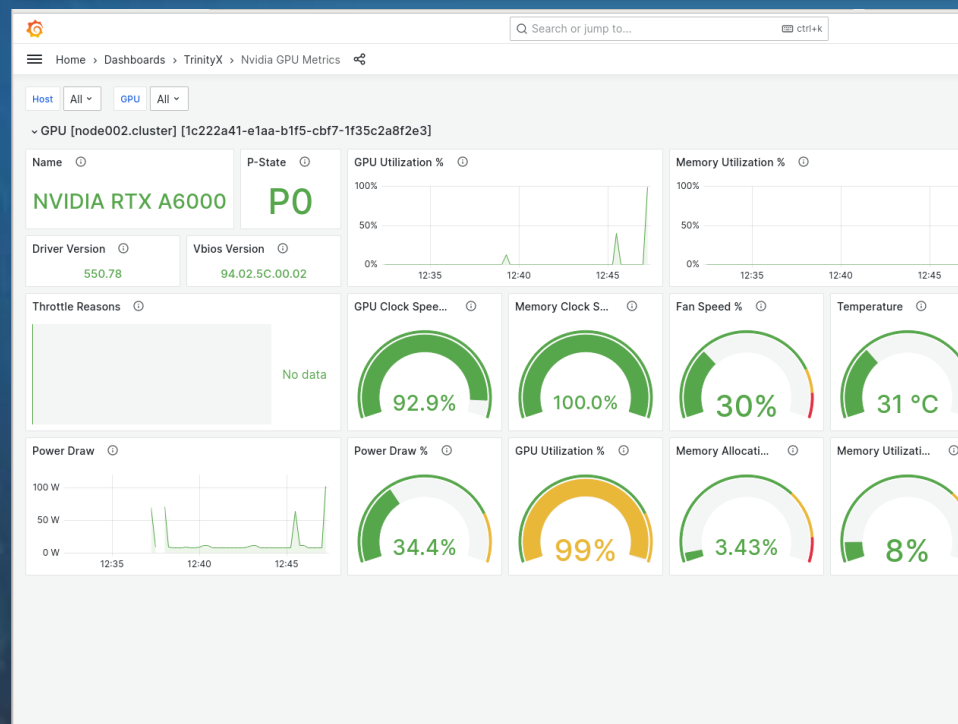


最初に Bright Cluster Manager を作った、ClusterVision社による、新設計のHPCクラスター構築・管理ツール。

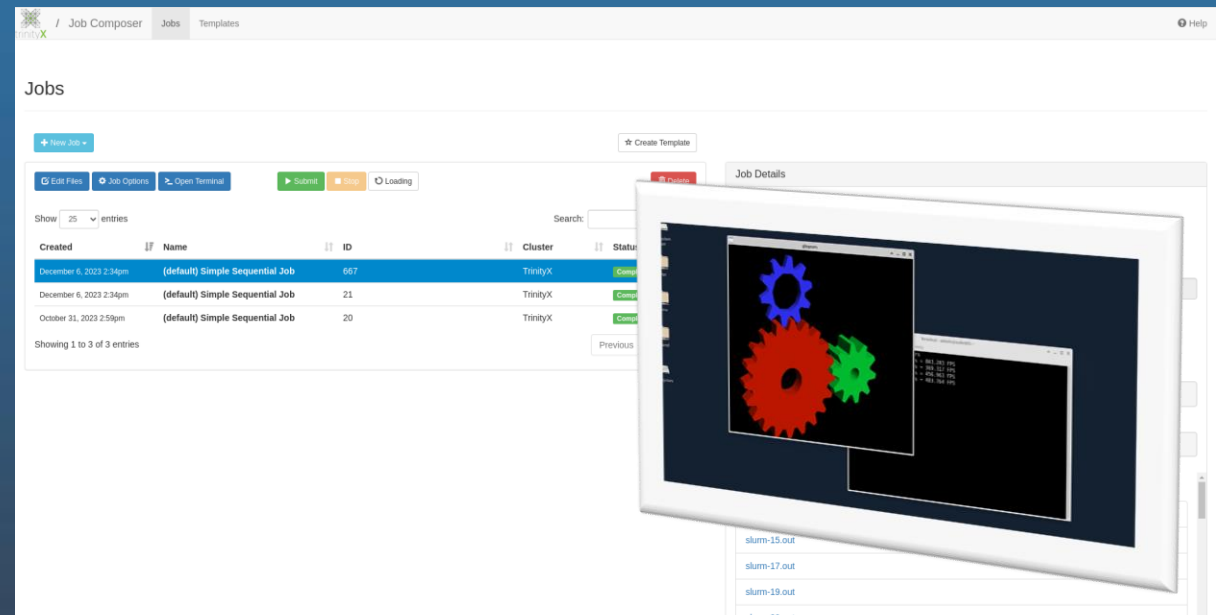
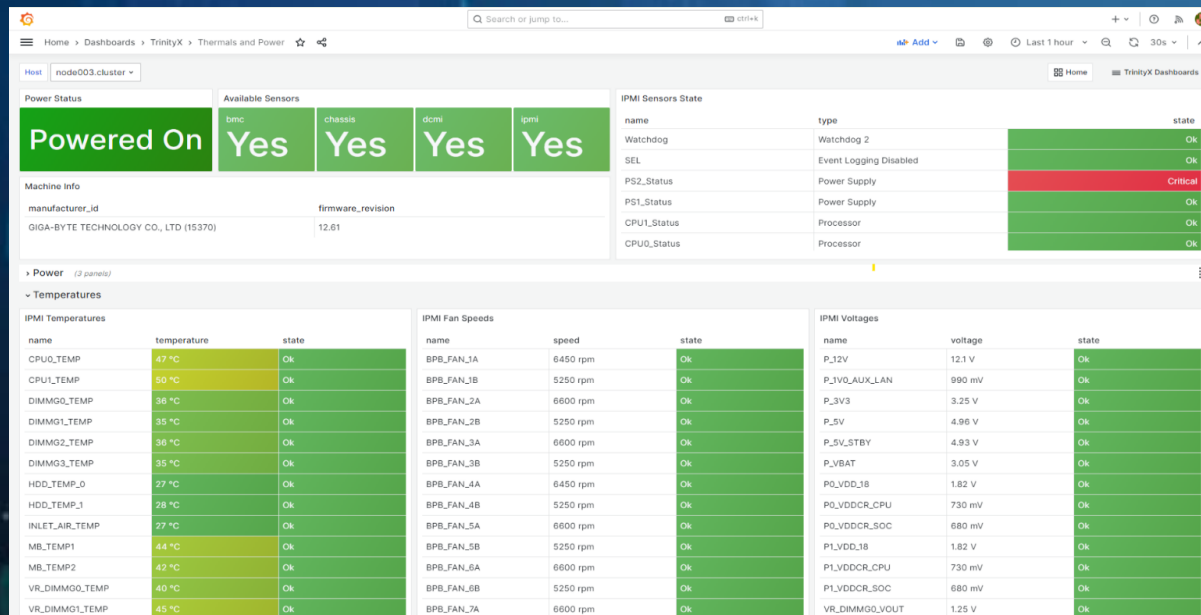
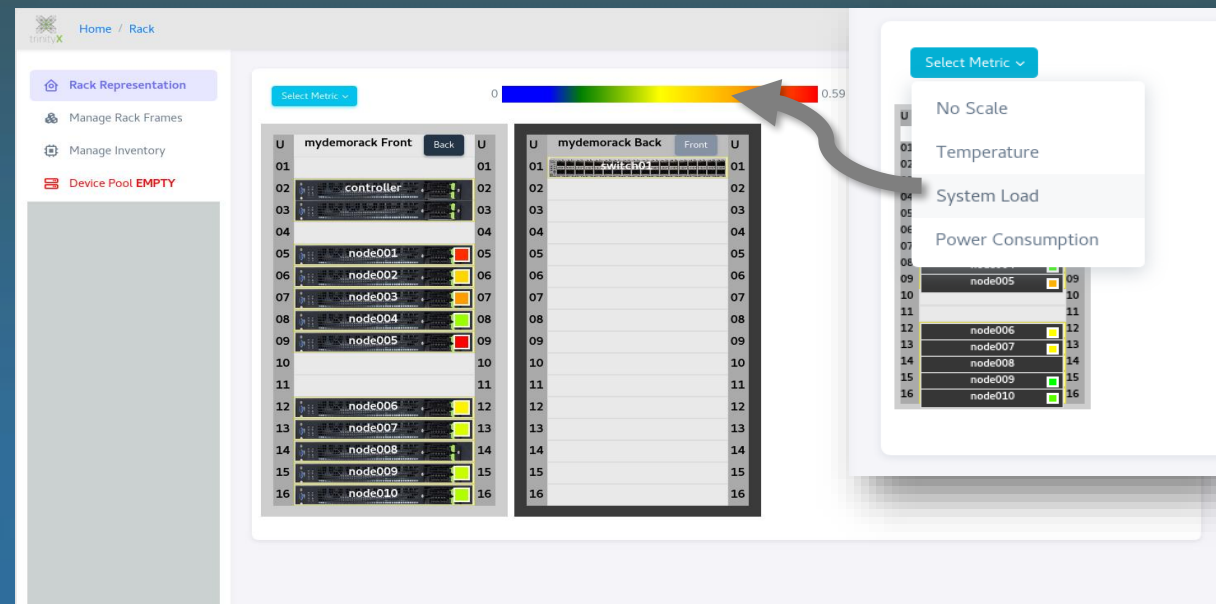
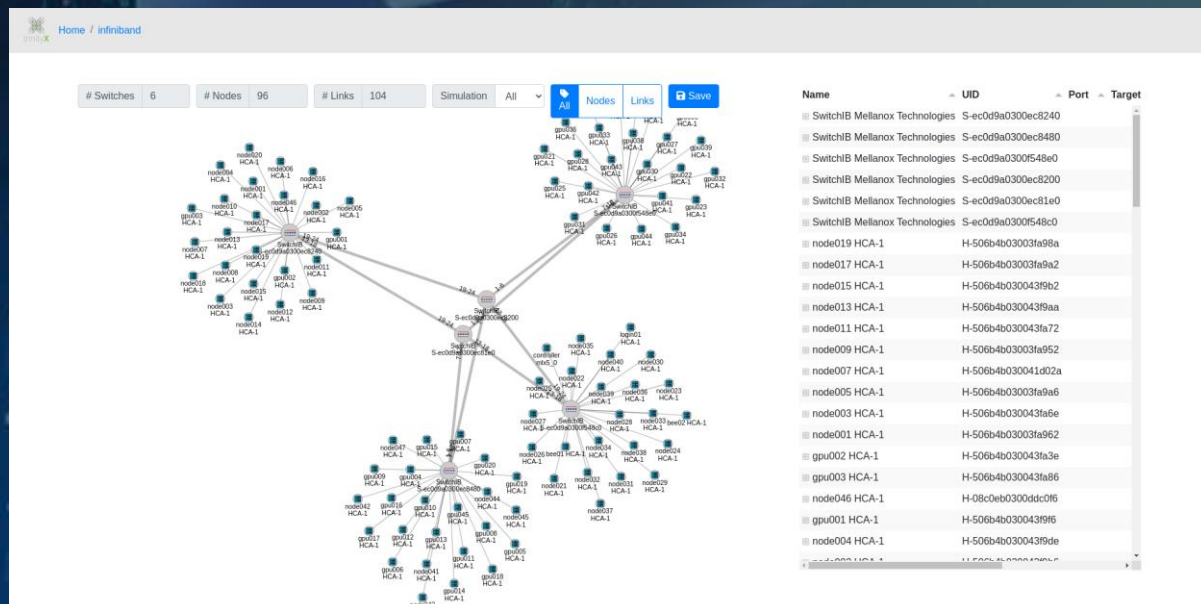
OpenOnDemand の上にユーザーポータルのみならず、管理ツールUIも統合。

充実したモニタリング、ジョブ実行状況分析に寄与する様々な管理画面。

計算ノードのOSプロビジョニングが超高速。
数千ノードでも一斉に展開。OS更新も円滑に。



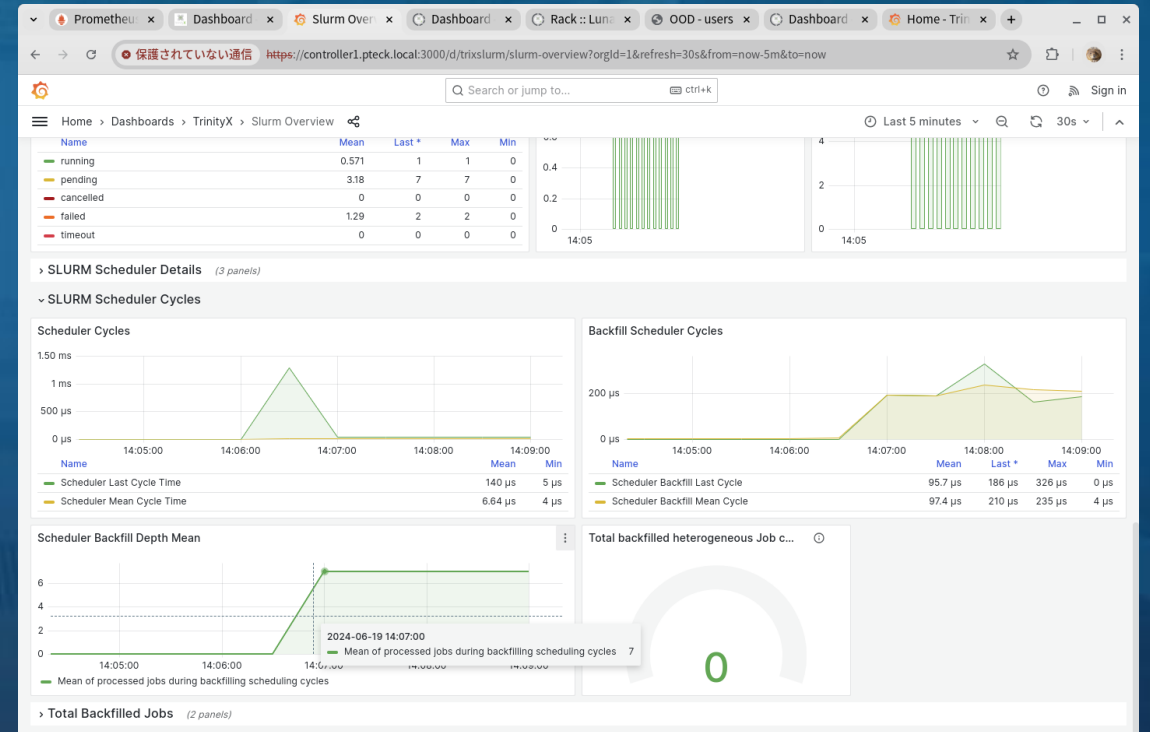
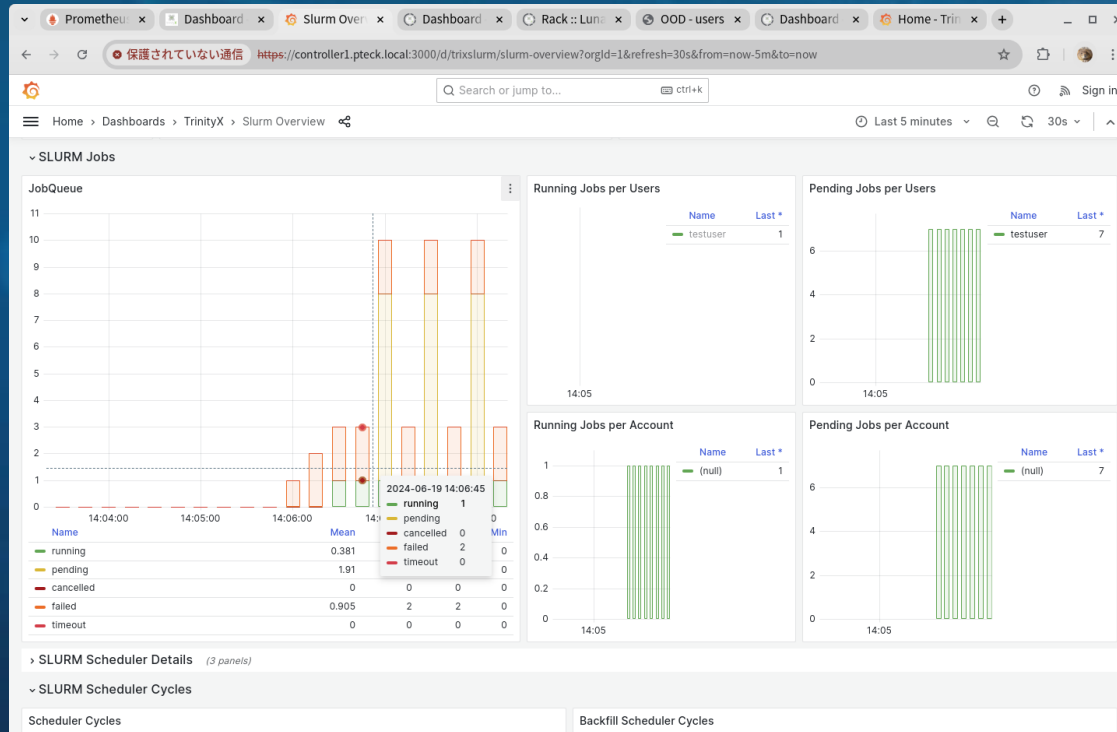
管理者にも利用者にも直感的なユーザーインターフェース



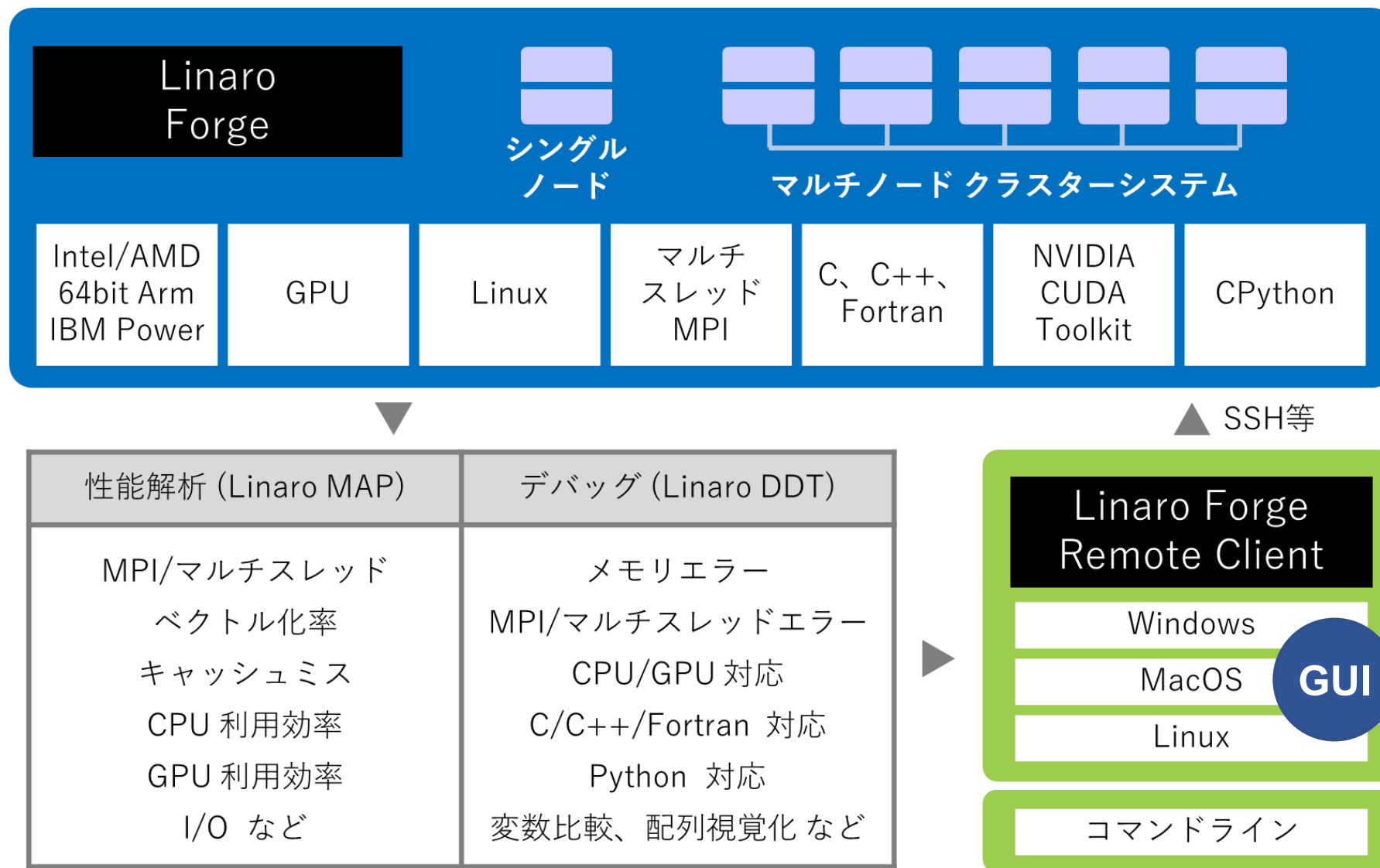
ジョブスケジューラの稼働状況の分析

ジョブの待ち時間等のデータを分析するダッシュボードがデフォルトで作成されており、利用状況に合わせたスケジューリング設定の調整だけでなく、将来的なリソース追加の計画策定にも寄与します。

slurm 設定の構築・検証ツールも付属しているため、GUIから必要な作業が完結します。



Linaro Forge によるアプリケーションの最適化



※一部の機能は Linaro Forge、Linaro Forge Ultimate のみで提供されます。

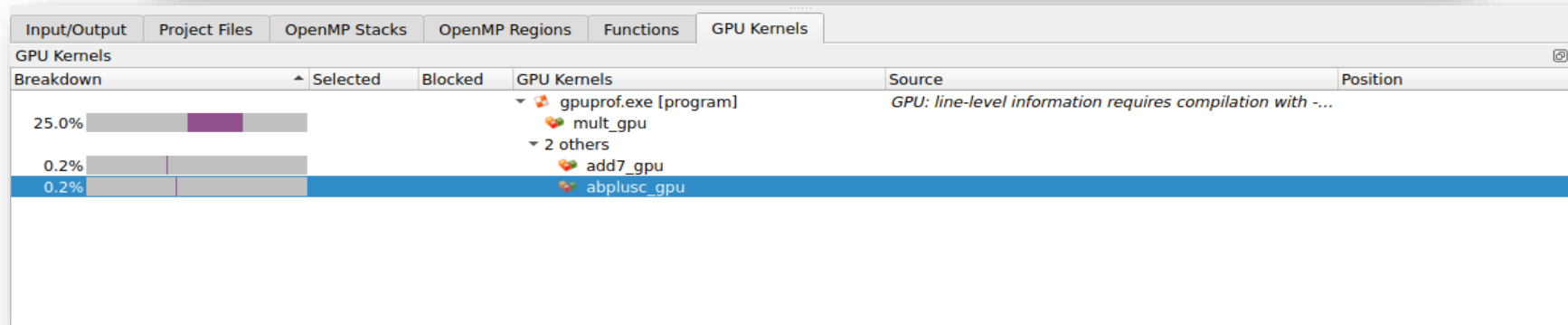
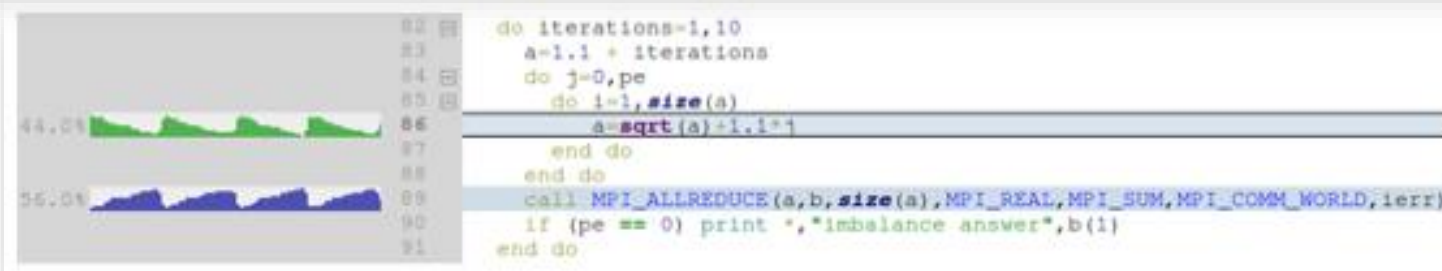
Linaro MAP

クラスターシステム向けの性能解析ツール。

MPI やマルチスレッド環境に対応。ソースコードの行レベルで性能問題を確認できます。メモリのプロファイリングでは、スレッドごとにアクセスパターンの分析ができます。また、消費電力の測定、削減の追求が可能となっています。

並列化に問題がある場合、通信時間や通信データ量をグラフ表示。ソースコードと照らし合わせながら、問題を容易に把握できます。

ARM, x86_64 などのCPUと、GPUとの組み合わせでのプロファイルが可能です。

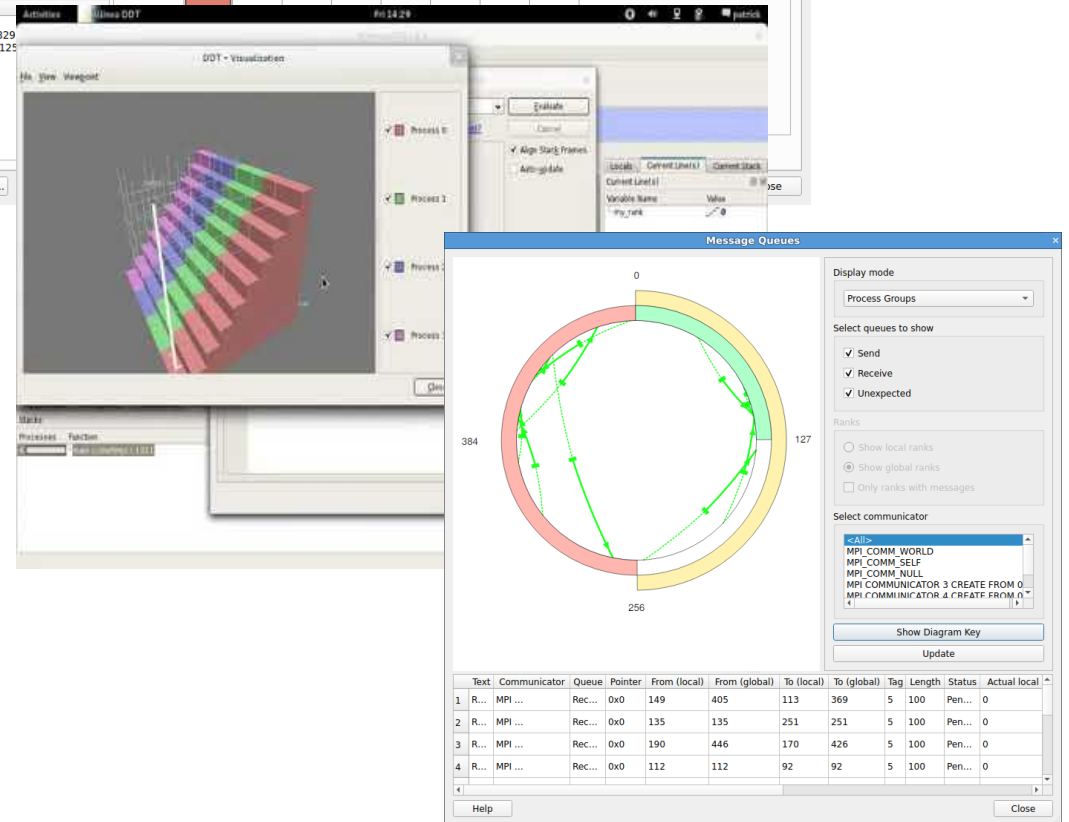
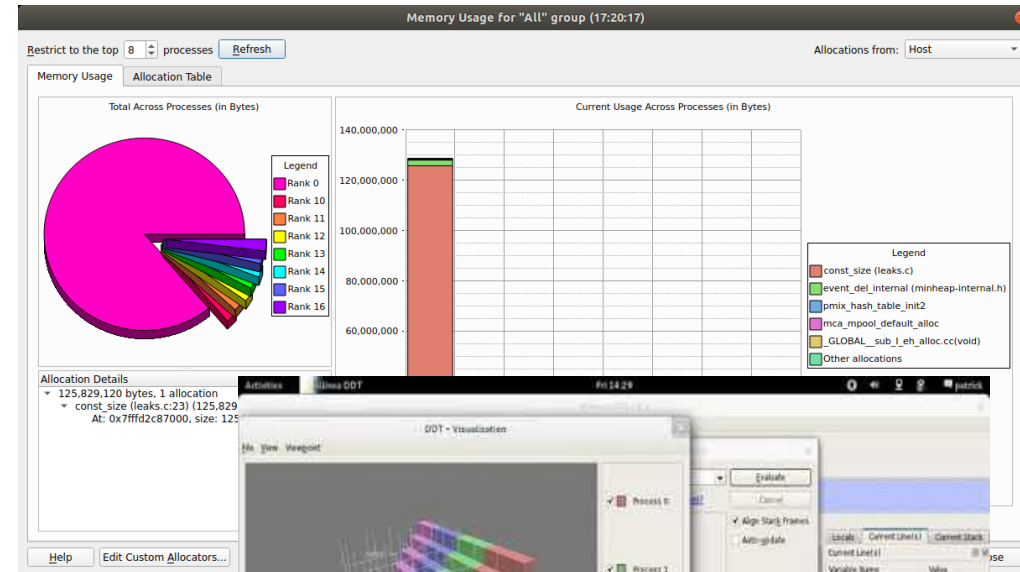
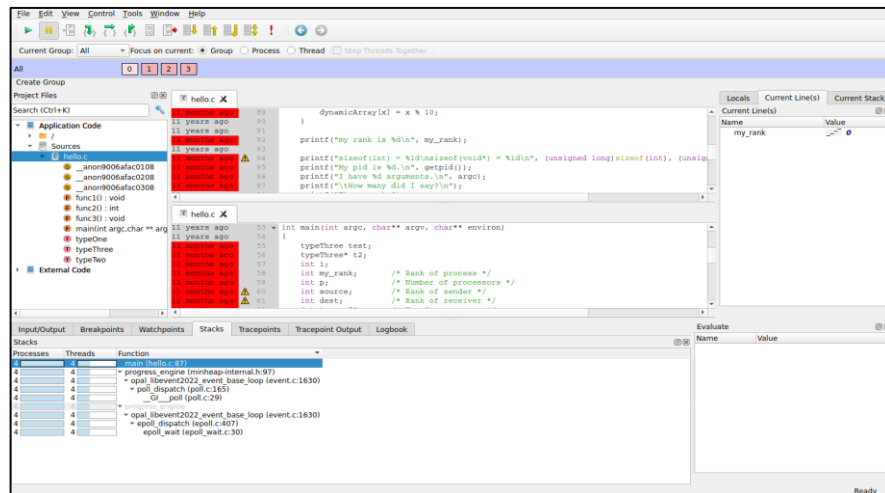


Linaro DDT

クラスターシステム向けのデバグガ。

- バグの自動検出
- メモリ使用量のグラフ化
- メモリアロケーションを追跡し不正アクセスを検知
- メモリリークによるリソース枯渇を防止
- 各プロセス・スレッドごとに対話型デバグ
- 変数の増減をグラフ化
- 多次元配列を立体描写
- メッセージキューの可視化デバグ

これらの機能を駆使し、並列計算における全プロセスの動作を容易に把握可能。





せっかくの資源を「制限いっぱいまで使い倒して絞り尽くす」ことに着目して、弊社のポートフォリオから2製品をご紹介しました。これがサステイナブルに寄与することを祈っております。

弊社ではOSSであっても、原則として開発元と Back to Back な有償サポート契約を軸にしたサービスを販売しています。
これにより、コード改変を含む責任ある対応が取れる他、開発者がサステイナブルに開発を続けられることで、HPCのエコシステムを守る意図があります。





Thank you

sales@pacificteck.com