
PacificTeck
HPC and Machine Learning Experts
It’s time to change.
ー変化が進化を加速するー
Supercomputing JAPAN! 2024
3月12日：タワーホール船堀

PacificTeck

HPC and Machine Learning Experts

世界中の最先端技術製品にフォーカス

- 日本を拠点にインドを含むアジア太平洋エリア(APAC)に製品を提供
- 英語 / 中国語 / 日本語のグロバールな言語での支援が可能
- ハードウェアに依存しないソフトウェアソリューションを中心に提供
- APAC最大のスパコンでの採用実績多数
- ストレージ/コンテナ/ジョブ管理のエキスパート

前回までの振り返り

1990年

- Windows 3.0 発売
- Microsoft Office
- GNU Hurd 開始
- WWW 提唱
- 応用数理学会発足

来日

1991年

- Linux Initial Release
- Python 0.9 公開
- PBS 開発開始
- MPI 仕様策定開始
- 数値風洞 開発開始
- LINPACK

1992年

- 第五世代プロジェクト終了
- RWCP 開始
- SINET運用開始
- CP-PACS 開始
- DEC Alpha 発表
- HPCwire 創刊

1993年

- LSF/GridEngine
- Beowulf 構想
- Intel Pentium 発売
- WindowsNT 3.1 発売

非常に多くの種がこのころに
撒かれている一方で、
Supercomputing の
在り方が変化し始めた時代。

取扱製品群

- ジョブ管理システム

ALTAIR

- HPC向けコンテナーシステム

S PRO　　S ENTERPRISE

- 並列ファイルシステム

BeeGFS　BeeOND　VAST

- オブジェクトストレージ

DATACORE　wasabi

- ストレージ連携ソリューション

iRODS　cunoFS unleashing object storage　Blocky for Veeam®

- クラスターマネジメントシステム

Bright Computing　PENGUIN SOLUTIONS

- I/O Profilingとプログラム開発者用ツール

oneAPI　Linaro Forge　ALTAIR

90年代前半の流行に乗せてご紹介します
ご笑覧ください。

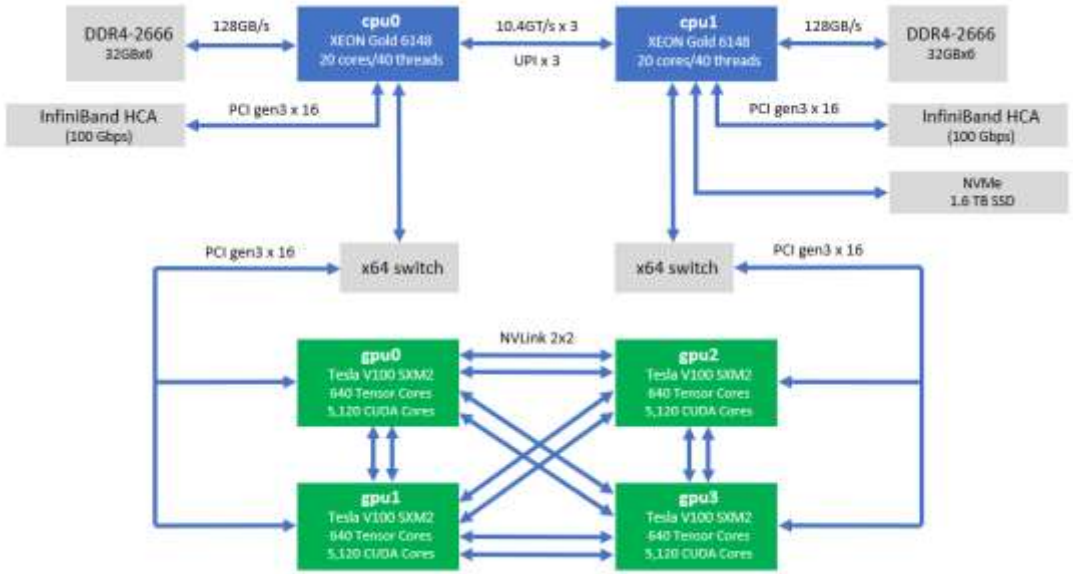
NUMAの配置

ALTAIR

ジョブがあふれる悲しい季節は キューから流れない夢を見る
転送中は遅いと言って最近接へ配置して 離れられない CPU & GPU

GPUの接続トポロジーをスケジューリングに反映可能。
「制約条件」として入れることは他のプロダクトでもできるが、設定や条件によってはスタックしてしまい、うまくジョブが流れないことも多く、スケジュールはかなり難しい。

NUMAをまたがせないパフォーマンスの最大化が可能。



grid engine

多数のジョブ中でリソースの予約を入れる機能の実現は、GPU の接続トポロジーと同様に極めて困難で、スケジューリングの難易度は一気に上がります。

安定したジョブの予約機能が必要であれば、Grid Engineの導入が最短距離のソリューションとなります。

コンテナ実行

PRO ENTERPRISE

ファイル一つだけでコンテナ実行へ in the cluster
使いまわせ my Image

Singularityとは、「コンテナエンジンとしてのSingularity」と「イメージの規格としてのSIF」、両方からなる概念です。そして、SIFを利用すること自体に際立った優位性があります。

高い可搬性

- シングルファイルで保存するため、持ち運びが用意。
- 通常のNASでの管理も可能。
- リポジトリに依存せず、電子署名による同一性を保証。

低フットプリント

- 内部はSquashFS。tar+gzipを直接マウントしているイメージ。
- オンメモリでアクセスでき、事前に展開する必要がない。メモリやCPU負荷も極めて小さい。

ハイパフォーマンス

- ファイルアクセスはローカルのオンメモリで行われ、共有ファイルシステムへの、メタデータアクセス負荷を極小化できる。通信負荷も下げられる。

AIのワークロードがPythonベースで開発。様々な環境がユーザーのホームディレクトリに展開されるケースが増えた結果、共有ファイルシステムの特にメタデータに高負荷を生み、大きな問題となっている。

例）PyTorchはimportするだけで1000回もファイルをオープンする。

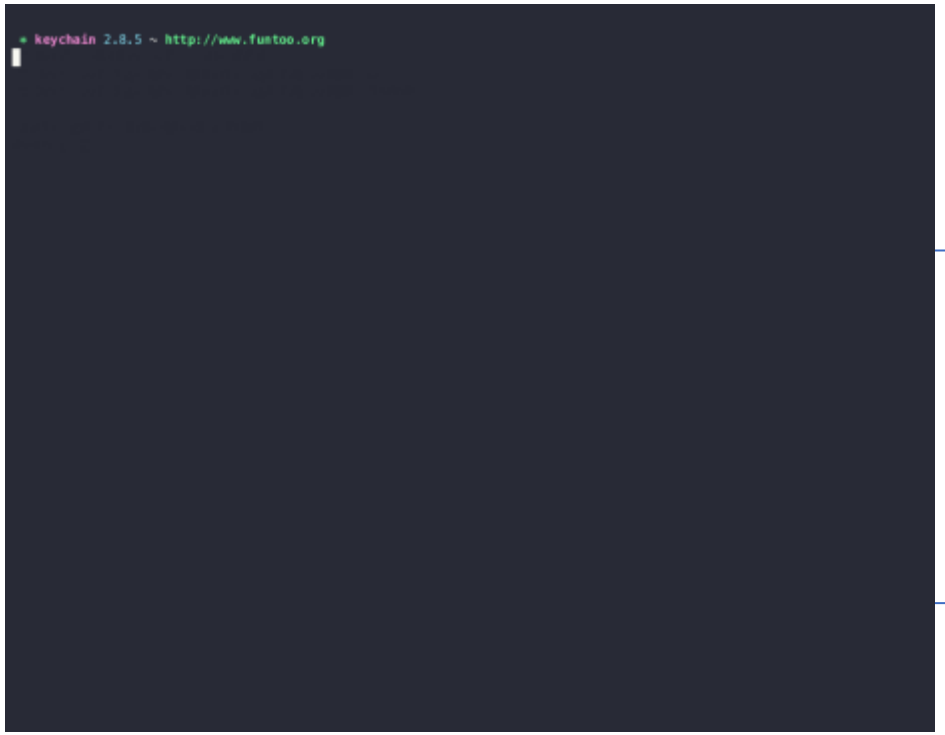
Dockerfileがあるだけで

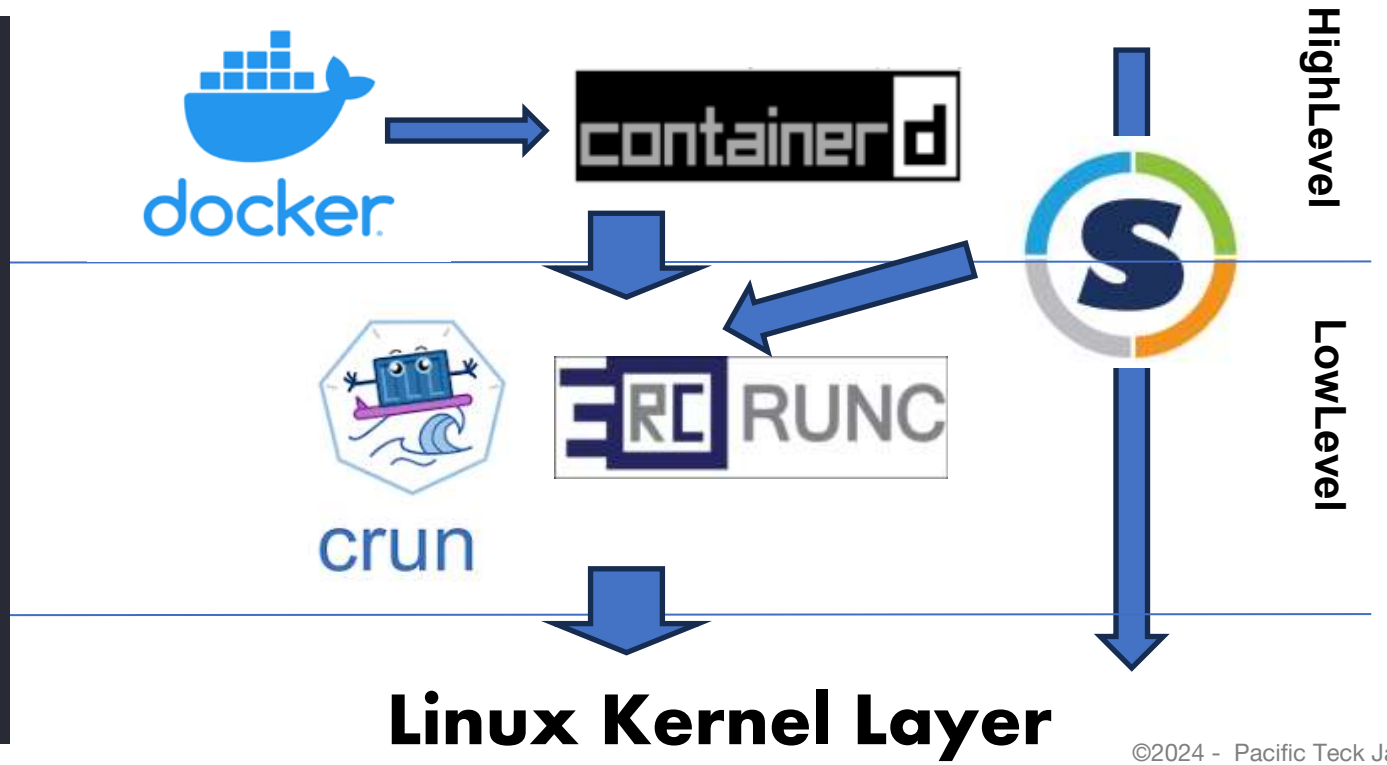
例えばdockerfileがあるだけでSIF Imageが作れること
互換性が大切なことを　気付かせてくれたね

SingularityCE/PROは昨年末 4.0 へメジャーバージョンアップ。最新は 4.1 です。

最も大きな変更は OCIモードの追加

- OCIイメージをSIFファイル形式で保存する OCI-SIF(Encapsulated OCI) image をサポート
- これに伴い、Dockerfile から直接、動作互換のあるOCI-SIF イメージ作成ができるように。
- SIF を OCI bundle として crun/runc から起動できるので、他のコンテナ実装との連携が容易に。





I/Oストールは突然に

何から調べればいいのか わからないままログは流れて
あの日 あの時 あのジョブで 応答がなくなったから

パフォーマンスを高める構成にいつでも変更できるフレキシビリティ。データ圧縮と重複排除、さらにSimilarity技術により、性能を維持したまま、物理容量を大幅に超える容量を実現。QoSにも対応。

ダッシュボードからリアルタイムのI/O挙動や利用状況、データ圧縮の状況まで完全にモニタリング。障害時の状況もチェック可能。

VAST

ALTAIR

mistral

breeze

Altair Mistral はジョブスケジューラと連携し、ストレージへのI/Oをジョブ毎に分析。全体挙動からボトルネック解析が可能。

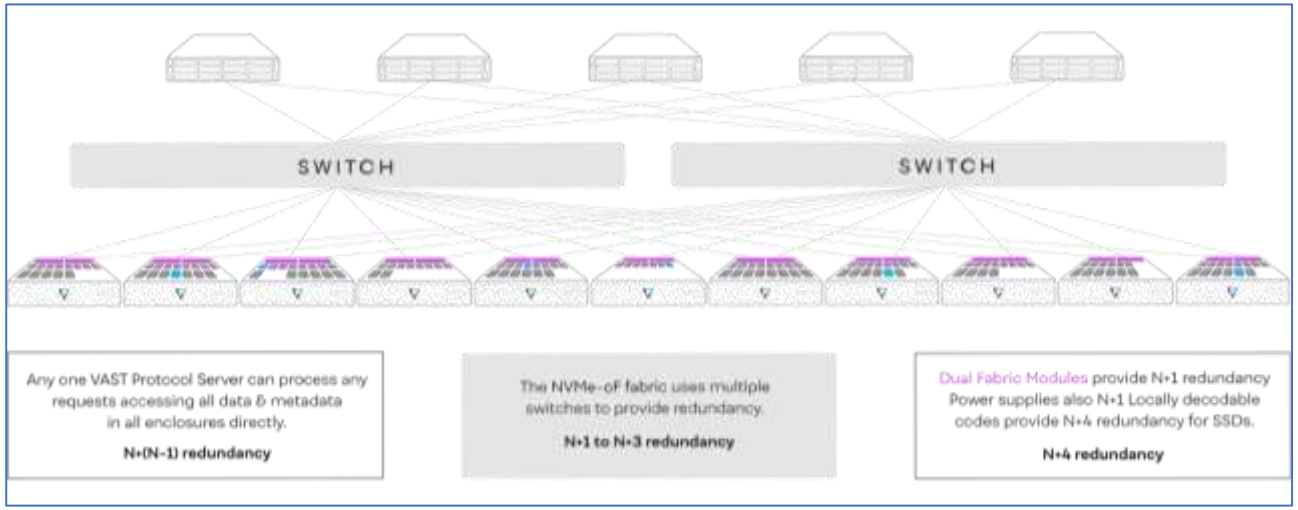
Altair Breeze はアプリケーションの I/O挙動をプロファイリング。良し悪しを採点し。チューニングとデバッグに有用な情報を提供。

壊れかけのRAID

何も読めない　何も書かせてくれない
　僕のデータが昔より　巨大になったからなのか

Share Everything の概念によりラックスケールの高可用性と、Erasure Coding（最大146/150）により、常に4本分のドライブフェイルに対する耐久性を維持。パリティのオーバーヘッドが最小化できるため、データ圧縮と併せ、物理容量を超える実効容量。

SCMを用いてQLCドライブへの書き込みを最適化。データ圧縮により、実際の書き込みデータ量を低減することで、長寿命化と速度向上を同時に実現。



オーバーヘッド無しでディレクトリごとに取得可能な1000階層までのスナップショット。隠しディレクトリとしてアクセス可能なため、ヒューマンエラーからの回復もユーザーで対応可能。

スナップショットからS3ストレージへのバックアップも、他製品を使うことなく実施可能。

ユーザークオータだけでなく、ディレクトリクオータに対応。QoSとあわせ、同一ネームスペース内に、用途別の性能や冗長性をも設定可能。

リージョン間で統合したネームスペースを構築し、複数のモードでの同期設定が可能。

おどるプロトコル

なんでもかんでもみんな データを保存しているよ
スパコンの外から SMBと　S3アクセス登場

独自の DASE™ により保存されたデータに対し、プロトコル変換を施すことで、同一のデータに対して多彩なプロトコルでのアクセスを実現。

独自プロトコルではなく、NFSを筆頭に標準的なプロトコルが使えるため、クライアント側に一切手を加えずに利用が可能。

プロトコル変換サーバーはステートレスに動作するため、台数比例でスケールする性能が得られるとともに、高可用性も同時に実現。

名前空間を分割し、IPアドレスベースのマルチテナントにも対応。

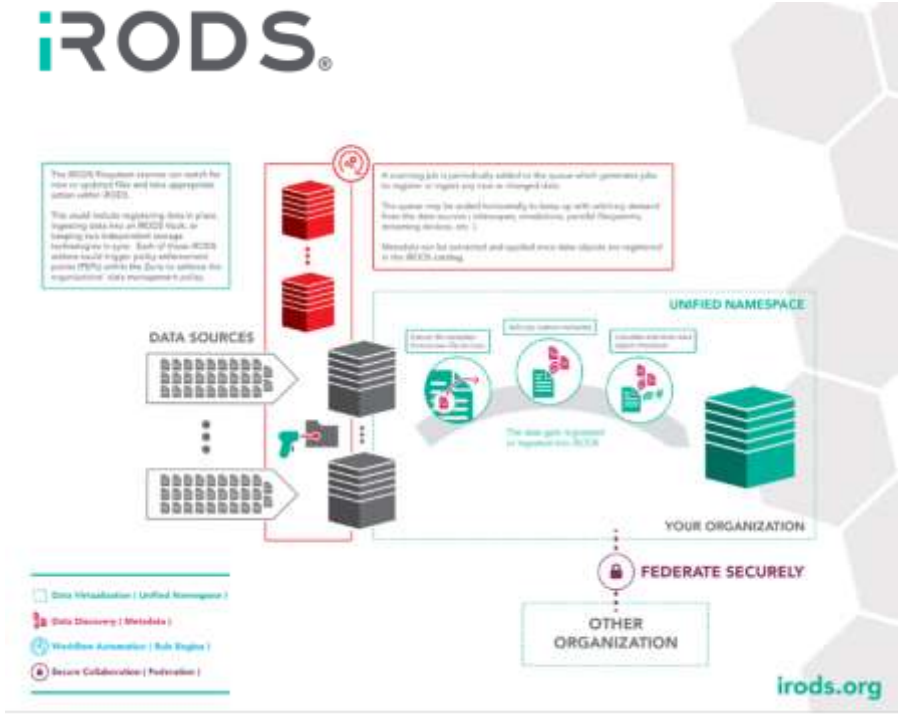
2カ月に一度のアップデートで、オンラインで機能強化。

世界中の誰より多く

世界中の誰より多く　データ吐き出してたから
目覚めてはじめて気づく　クオータあふれに

Swarm は世界初のオブジェクトストレージと言える存在。
運用が極めて平易で容量の利用効率が高く、独自の HTTPSでの転送のほか、S3でもアクセスが可能。またオプション製品でNFSやSMBでのエクスポートにも対応。

iRODS は複数のストレージを束ねた仮想ファイルシステム空間とし、ユーザー定義のメタデータを用いてデータハンドリングや自動処理を実現するミドルウエア。インテリジェントなストレージを構築する。
データの出し入れは独自プロトコルだが、NFSやS3への変換も可能。



もうFTPなんてしない

旧来のスパコンの付属物としてのストレージではなく、多彩な入出力を可能とした多機能ストレージを中心に据え、スパコンはデータ処理サブシステムのような存在になるのではないか。

クライアントがないと　何もできない訳じゃないと
コマンドを叩いたけど　鍵のありかがわからない

平成

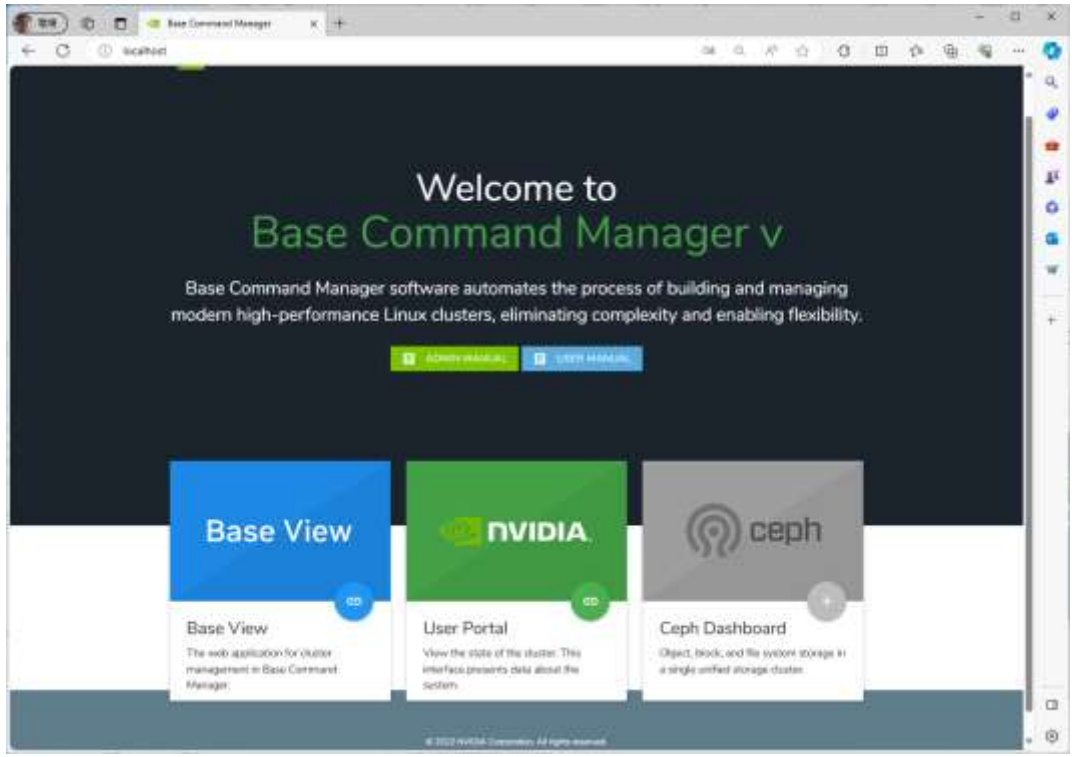
令和

Designed by slidesgo and macrovector / Freepik

サーバーとGPUと私

サーバーとGPUと私　大事な計算のため
まいにち　安定運用したいから

NVIDIA Base Command Manager(旧Bright Cluster Manager),
PENGUIN Scyld Clusterware はどちらも、クラスターの構成管理、
構築、モニタリングを最小限の工数でワンストップに実現。
数台～数千台以上にスケールするHPC環境に加え、クラウドバースト環境の
構築にも対応。

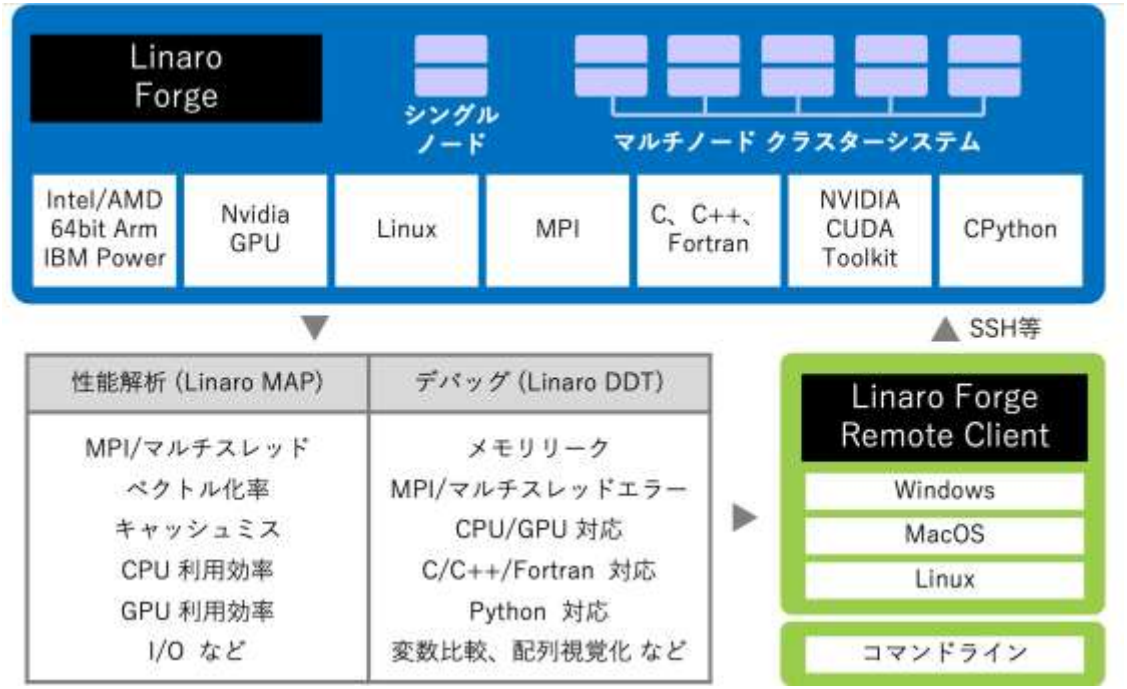


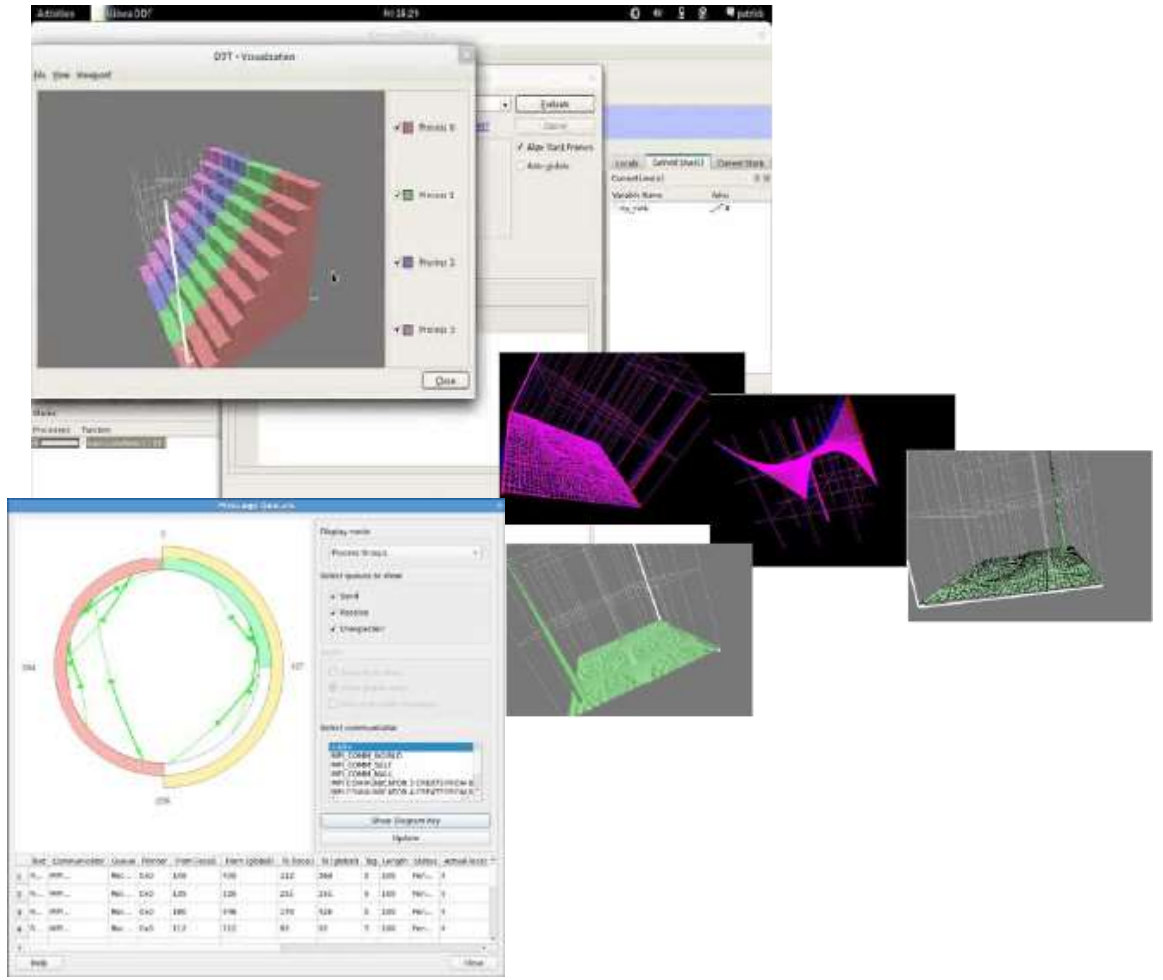
もうチューニングしないなんて

ほらバイナリもつくれたもんね　だけどあまり速くない
-fast で作ったのなら文句も　思い切り言えたのに

Linaro Forge

Allinea DDT ⇒ Arm Forge ⇒ Linaro Forge と変遷。高い評価を受けて来た歴史あるデバッガとプロファイラ。
CPU（x86_64, Arm をはじめとした主要アーキテクチャに広く対応）
GPU（NVIDIA, AMD）、MPIによる並列でも直感的にメモリの中身やメッセージを可視化してデバッグ可能。





It’s time to change Supercomputing.

高校物理の教科書を見ておどろくのは、私たちの頃と内容が変わってないことです。
20世紀以降の物理学の飛躍的な進歩はほとんどおまけのような扱いです。
一方、高校生物の教科書を見てみると、自分が昔の人であることを感じざるをえません。
まるで様変わりしているからです。

橋本省二「思いは伝わっているか」日本物理学会誌 No.2 VOL.79 2024

Now that we have changed the world, it's time to change America.

Bill Clinton 1992 Acceptance Speech
1992年 日本流行語大賞特別賞 アメリカ合衆国大使館

もう既にSupercomputingは、その成果で十分世界を変えてきました。
そろそろSupercomputing自身の在り方を考えるタイミングではないでしょうか。
あの90年代初頭のように。その一助を担えましたら幸いです。

Thank you

sales@pacificteck.com