複数の ARM マシンを集約するハードウェア仮想化レイヤ
- Pilevisor：マシンの資源を集約するハイパーバイザ -

1　背景

計算資源の処理性能向上方法として，一つのマシンのスペックを向上させるスケールアップと，マシンの台数を増やしコンピュータクラスタを作り，全体の性能を向上させるスケールアウトが挙げられる．

特に，スケールアウトは導入のコストの安さやシステム構成の柔軟さがメリットである．しかし，Windows や Linux などの汎用 OS や，一般的な共有メモリ向けアプリケーションはクラスタ上では動作しない．

クラスタ上の資源を利活用するには，専用のソフトウェアを使う必要があり，用途が限定される．また，クラスタ上で高速なプログラムを書くときには，OpenMPI などメッセージパッシング型プログラミングのためのミドルウェアのインタフェースに従う複雑なプログラミングを習得する必要があり，普段 Python などの一般的に普及しているプログラミング言語を利用しているユーザにとっては敷居が高い．

このように，我々のような普段汎用コンピュータを利用している一般ユーザがスケールアウトの恩恵を受ける簡単な手法が無いという現状がある．

2　目的

本プロジェクトでは，一般ユーザのコンピュータクラスタとのギャップを埋め，誰でもクラスタ環境を使いやすくするために，クラスタ上の資源を集約した単一システムイメージを構築する．

単一システムイメージを構築することによって，

- スケールアウトによる既存の汎用 OS・共有メモリアプリケーションの高速化
- 資源の単一化によるプロセスなどの資源管理の簡便化

を狙うことができる．

さらに，単一システムイメージを，ハイパーバイザによって仮想マシンとして構築するというアプローチを採用した．これにより，上で動作する汎用 OS やアプリケーションはマシンが分散していることを意識せず，クラスタの資源を利活用できる（図 1）．

3　開発の内容

本プロジェクトでは，今後急速に普及していくであろう強力かつ低消費電力な ARM アーキテクチャのマシンを対象に，複数の物理マシンと Gigabit Ethernet によって構成されるクラスタに跨って動作するハイパーバイザである “Pilevisor” を実装した．実装されるハイパーバイザはクラスタの資源を集約するためのハードウェア仮想化

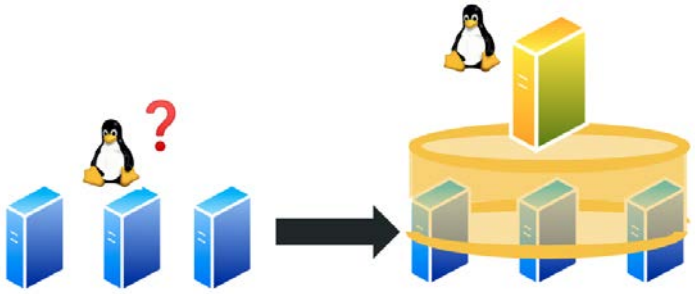

図 1: 本プロジェクトのアプローチ

レイヤとして動作し，クラスタの資源を集約した単一のマシンを仮想的に構築する．仮想マシンは，物理マシンから提供された資源（CPU，メモリ）を全て集約したもので構成される．例えば，4CPU, RAM 4GB を仮想マシンに提供する物理マシンが 3 つ集まると，12CPU, RAM 12GB の仮想マシンができる．また，デバイスに関しては集約されず，1 つのものを全マシンで共有して使用する．仮想マシンは，これらの資源が分散していることを意識せず，まるで単一マシン内の資源として透過的にアクセスすることができる．システムの概要図を図 2 に示す．

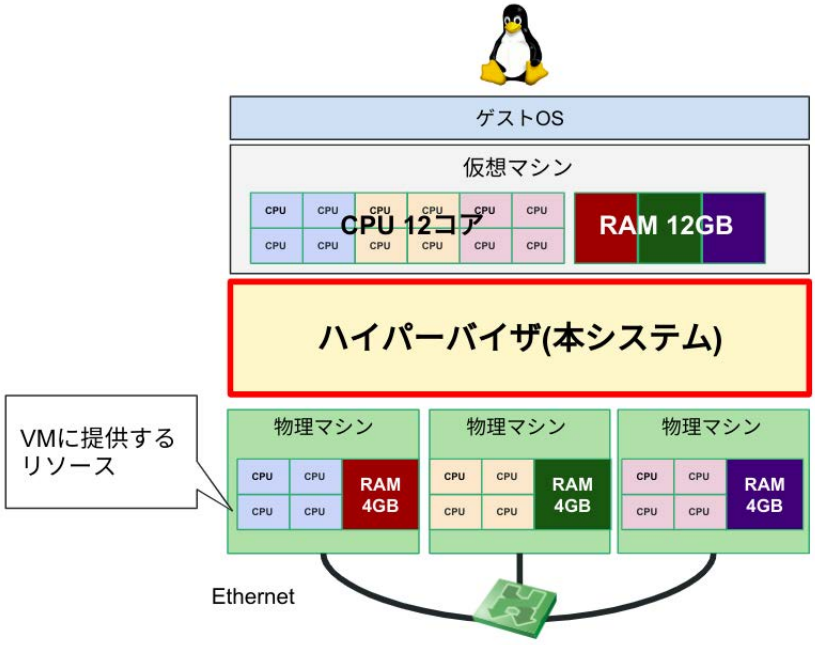

図 2: システムの概要図．赤線で囲まれた箇所を実装した．

以上のハイパーバイザを，C 言語と ARM アセンブリを用いフルスクラッチで実装した．仮想マシン上で動作させる OS に Linux を想定し，ハイパーバイザをエミュレータ環境である QEMU virt ボードと実機の Raspberry Pi 4B 上に実装した．実装したハイパーバイザの大まかなコンポーネントは以下の通りである．

- ハイパーバイザカーネル

メモリなどの物理マシンの資源管理や仮想マシンの管理など，システムの中核を担う．

- デバイスドライバ
  Ethernet でノード間で通信するための NIC ドライバやコンソール出力のための UART ドライバ，割り込みコントローラなどを実装した．
- ノード間通信
  仮想マシンが，分散したリソースにアクセスできるようにするために，クラスタのノード間で通信を行うことで，適切なエミュレーションを透過的に行っている．動作概要を図 3 に示す．ノード間通信のプロトコルは，独自に設計している．
- 仮想 CPU
  分散している CPU が単一システム内にあるように見せかけるための仮想 CPU を実装した．CPU 間の割り込みや起床はノード間通信を介して行われている．
- 仮想共有メモリ
  分散したメモリを透過的に単一のメモリ空間として扱うために，分散共有メモリを実装した．他ノードへのメモリアクセスも透過的に行うことができる．
- 仮想デバイス
  仮想割り込みコントローラや仮想タイマなどを仮想マシンに提供している．

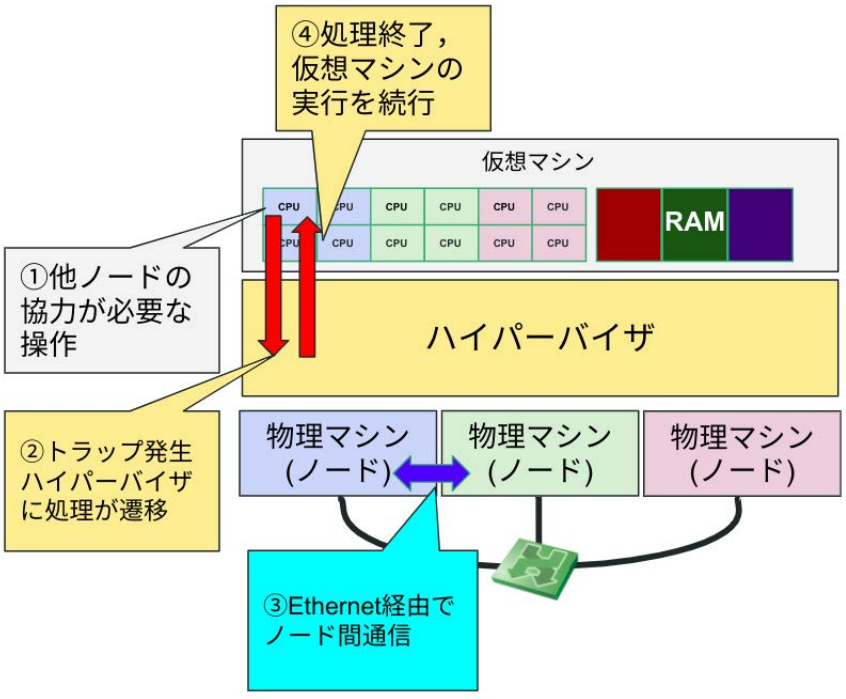

図 3: 他ノードの協力が必要なときの Pilevisor の動作概要

本プロジェクトでは，QEMU virt ボードのマシンを集約した仮想マシン上で無改変の Linux と共有メモリ向けアプリケーションを動作させることに成功した．Linux が，本来分散された資源を単一のマシンとして認識できていることを確認した．

4　従来の技術との相違

本プロジェクトはハイパーバイザで単一システムイメージを提供するという仕組みを ARM アーキテクチャに適用した世界初のプロジェクトである．さらに，単一システムイメージを構築する Type-I ハイパーバイザの中で，実装が全てオープンソースで公開されているのは，本プロジェクトのみである．

5　期待される効果

本プロジェクトは，ハイパーバイザのレイヤで単一システムイメージの構築をすることによって，仮想マシン上で動作する OS やアプリケーションの種類を限定せず，一般ユーザがスケールアウトの恩恵を簡単に享受できる．これにより，安価に高性能なマシンを構築できることが期待できる．さらに，クラスタノードの増減だけで簡単にマシンの性能の向上・低下が実現できるため，柔軟かつ扱いやすいシステムを構成することができる．

そして，全ての実装はオープンソースとして公開されているため，後続の分散ハイパーバイザ分野の研究の礎となり得ることが期待される．

6　普及の見通し

本プロジェクトは既に実装が公開されているが，ドキュメントを整備し，一般ユーザがもっと気軽に使えるようなプラットフォームとして普及していくようにする予定である．将来，RDMA 技術や 100G NIC などの汎用的かつ高性能なインターコネクトの普及で本システムの性能が向上し，誰でも安価かつ強力なクラスタを簡便に扱える時代が来ると考えている．

7　クリエータ名（所属）

- 飯田 圭祐（慶應義塾大学環境情報学部環境情報学科）
- 柚山 大哉（慶應義塾大学環境情報学部環境情報学科）

（参考）関連 URL

- ソースコード：`https://github.com/k-mrm/Pilevisor`