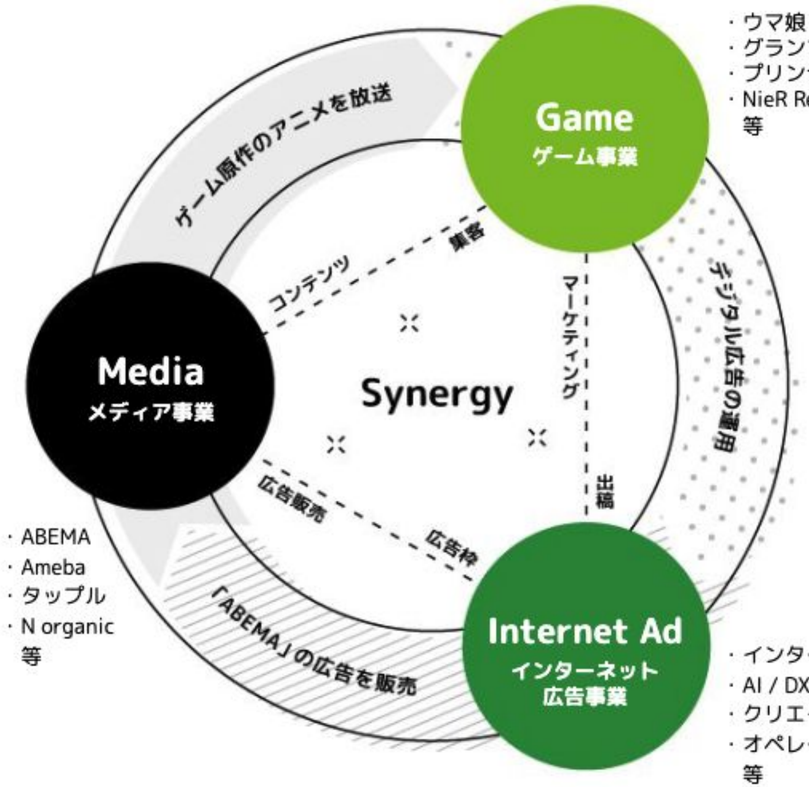
コンピュートノードのネットワーク高速化

Akira Kamio

サイバーエージェントについて

サイバーエージェントのビジネスモデル

メディア事業、インターネット広告事業、ゲーム事業を中心に事業を展開。
各事業とも技術力、運用力を競争力に事業を拡大し、事業シナジーを生み出しています。



・ウマ娘 プリティーダービー
・グランブルーファンタジー
・プリンセスコネクト！Re:Dive
・NieR Re［in］carnation
等

・ABEMA
・Ameba
・タップル
・N organic
等

・インターネット広告事業
・AI / DX 事業
・クリエイティブ事業
・オペレーション事業
等

経済的価値

・売上
・利益
・キャッシュ

社会的価値

・新しい産業の創出
・雇用の促進
・新たな情報インフラ
・IT人材育成
・エンターテインメントの提供

CA のクラウドについて

- **プライベートクラウドとしてサービスに提供**
  - IaaS
  - SaaS
    - DB
    - Kubernetes
    - GitHub Actions Runner
    - etc…

- **規模感**
  - HV: 1000+
  - VM: 5000+

- **Diskless Hypervisor**
- **AKE (Astro Kubernetes Engine)**

Motivation

- **仮想化にはオーバーヘッドが存在する**
  - 仮想化したサーバは物理環境に比べて、**仮想化レイヤの分だけ性能が劣化**してしまう
  - KVM という仕組みによって CPU やメモリはほぼオーバーヘッドなく利用できるように
  - ディスク IO やネットワーク IO はまだまだ改善の余地がある

- **仮想化したネットワークは遅い**
  - コンピュートノード上の様々なレイヤがボトルネックになる

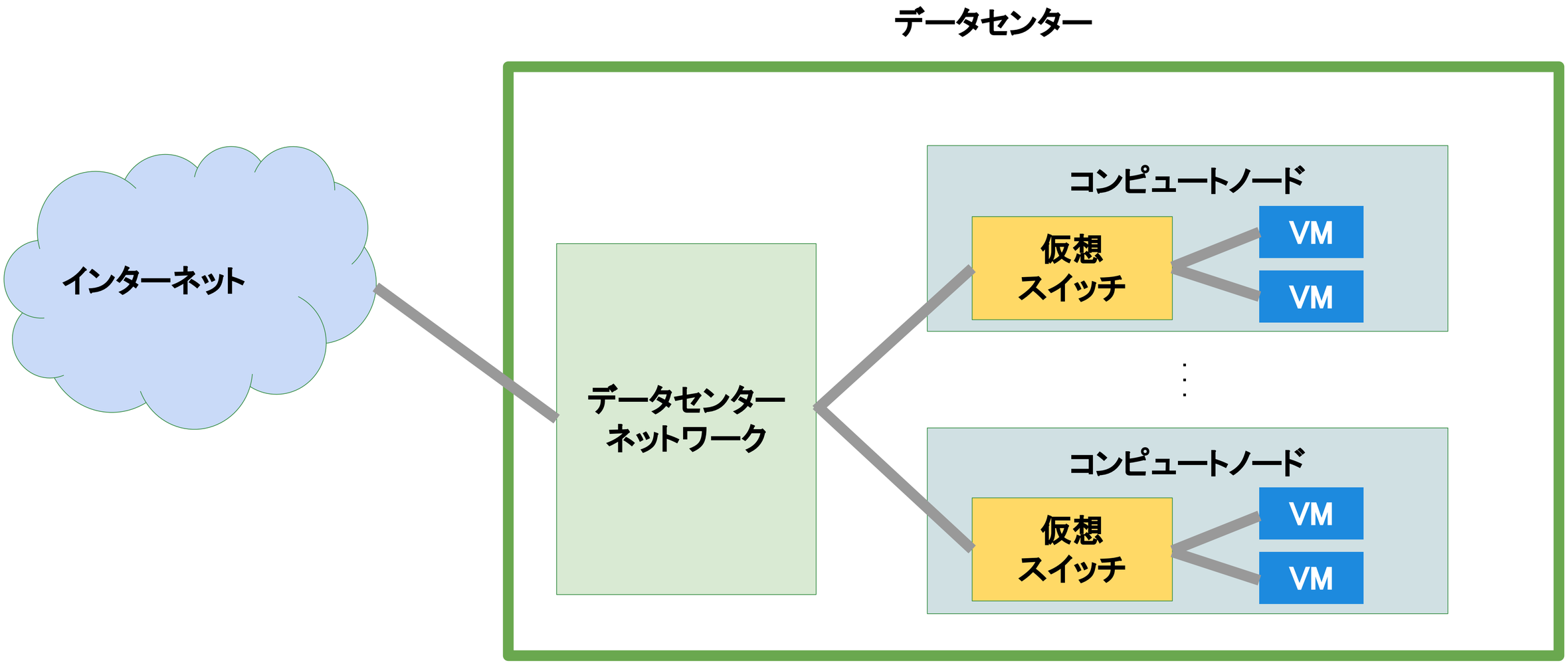
- **VM のネットワークを高速化したい**
  - **より高密度な仮想化環境を安定・低遅延に提供**することができる
    - ネットワーク負荷の高いサービスのサーバの仮想化
    - 低レイテンシな VM ネットワーク環境の提供

Contents

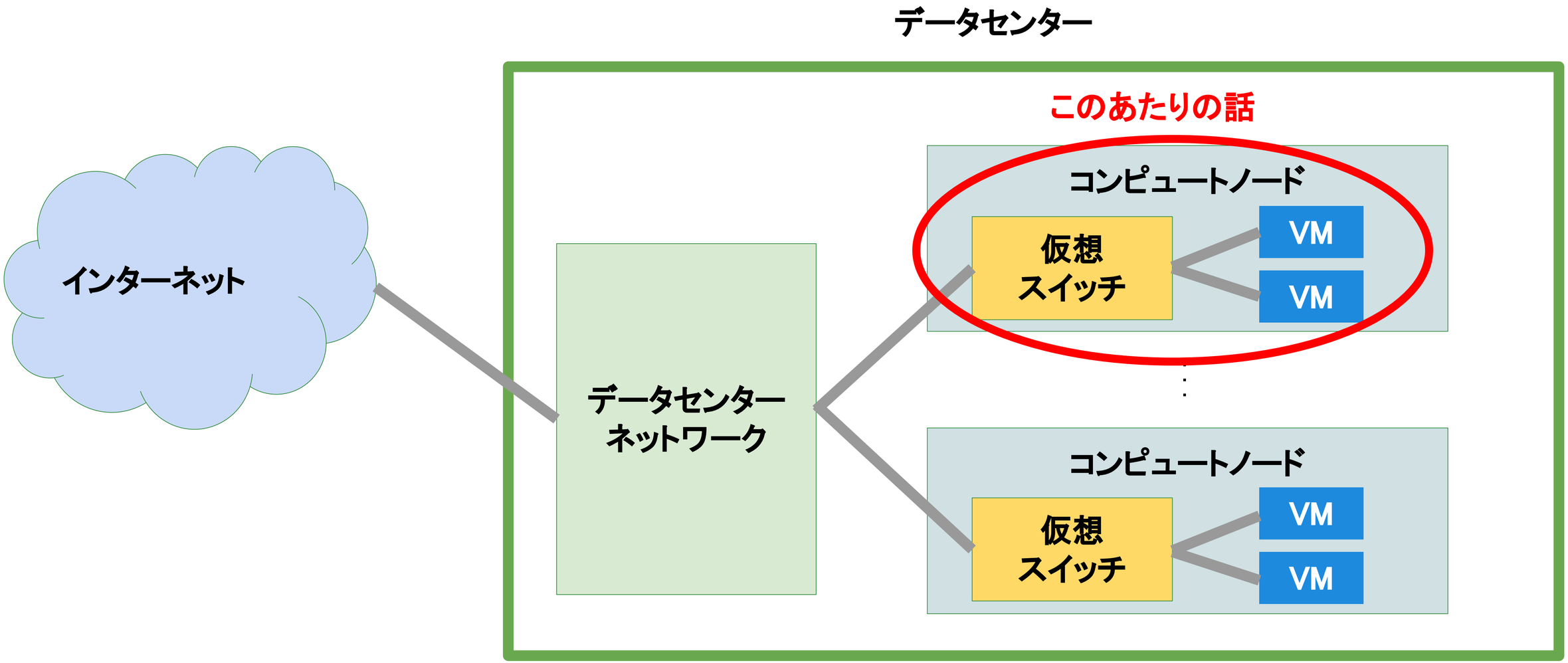
1

HV ネットワーキング

HV ネットワーキング？


データセンター
インターネット
データセンター
ネットワーク
コンピュートノード
仮想
スイッチ
VM
VM
コンピュートノード
仮想
スイッチ
VM
VM

HV ネットワーキング？


データセンター
このあたりの話
インターネット
データセンター
ネットワーク
コンピュートノード
仮想
スイッチ
VM
VM
コンピュートノード
仮想
スイッチ
VM
VM

HV ネットワーキング

ここでは下記の構成について述べる

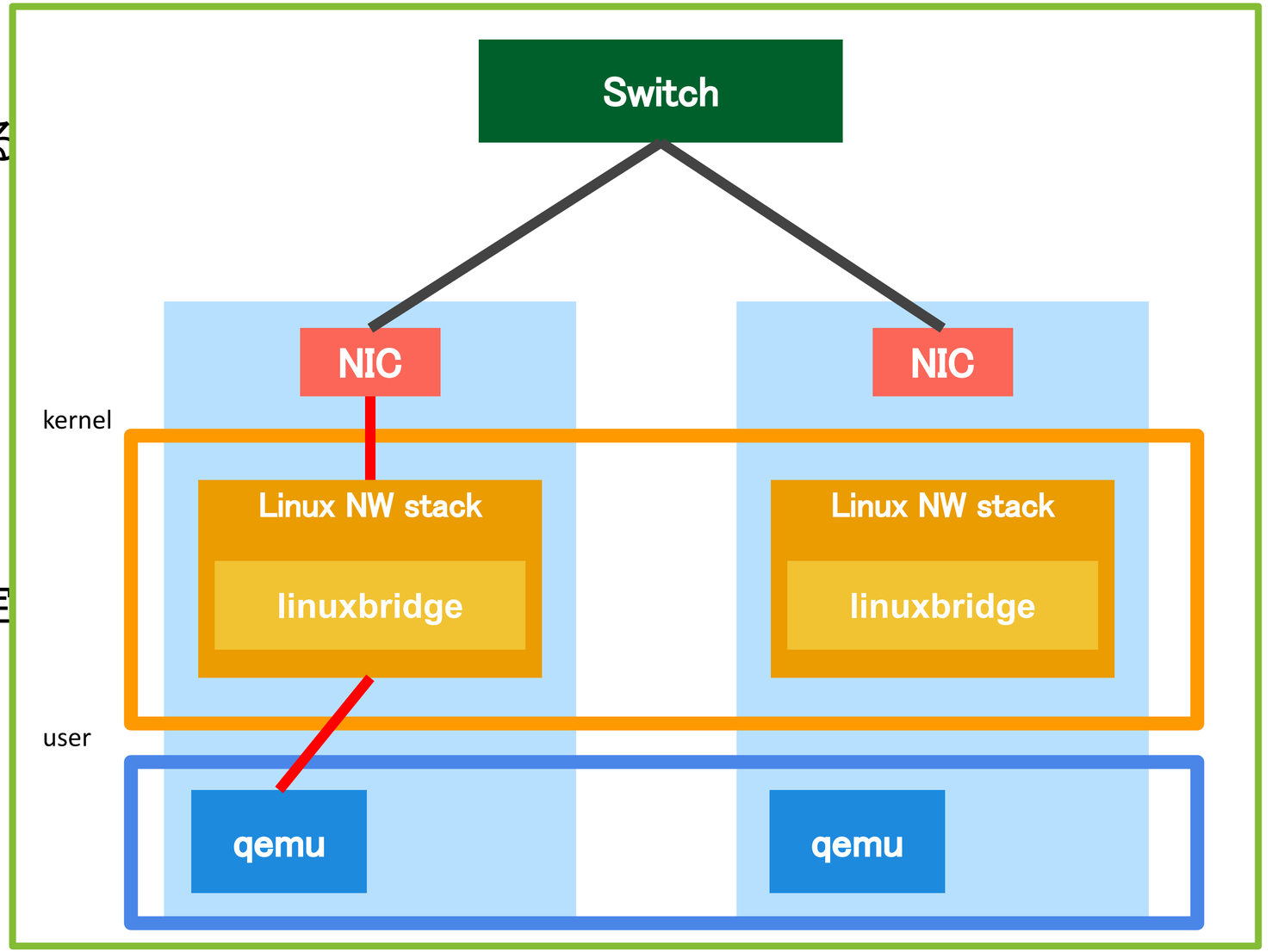
- **一般の家庭でもよく使われる**
  - linuxbridge
  - OpenvSwitch
- **高速化手法**
  - OpenvSwitch + **DPDK** + vhost-user
  - OpenvSwitch **HW Offload**
  - OpenvSwitch + **vDPA Kernel Framework**

linuxbridge

- **特徴**
  - Linux Kernel の機能として実装されている
  - 仮想 L2 スイッチが実現できる
- **性能**
  - ～20Gbps
- **Pros.**
  - 特別なパッケージを必要としない
  - kernel 空間で動作するため比較的高速
  - 実装が枯れている（バグが少ない／安定している）
- **Cons.**
  - L3 が喋れない

linuxbridge

- **特徴**
  - Linux Kernel の機能として実装されている
  - 仮想 L2 スイッチが実現できる
- **性能**
  - ～20Gbps
- **Pros.**
  - 特別なパッケージを必要としない
  - kernel 空間で動作するため比較的高速
  - 実装が枯れている（バグが少ない／安定
- **Cons.**
  - L3 が喋れない

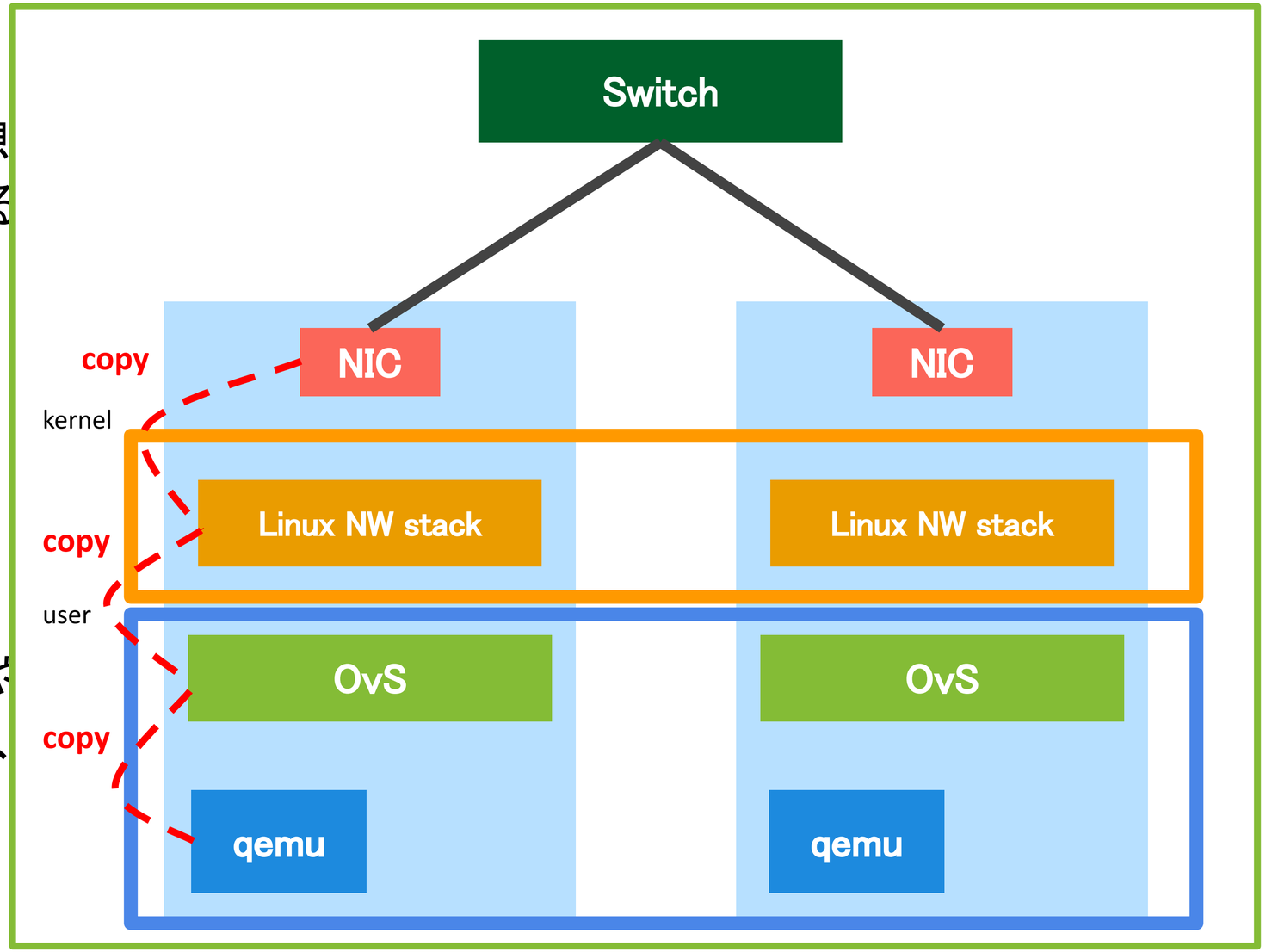


OpenvSwitch

- **特徴**
  - オープンソースの OpenFlow 実装の仮想スイッチ
  - linuxbridge と異なり user 空間で動作する
- **性能**
  - ～20Gbps
- **Pros.**
  - 手軽に構築できる
- **Cons.**
  - メモリコピーが多く発生する
    - softirq が多い -> CPU 使用率が上がりやすい
    - user 空間で実行されるプロセスのため、タスクスケジューリングの影響を受けやすい
    - **レイテンシが安定しない**

OpenvSwitch

- **特徴**
  - オープンソースの OpenFlow 実装の仮想
  - linuxbridge と異なり user 空間で動作する
- **性能**
  - ～20Gbps
- **Pros.**
  - 手軽に構築できる
- **Cons.**
  - メモリコピーが多く発生する
    - softirq が多い -> CPU 使用率が上がりや
    - user 空間で実行されるプロセスのため、
    - **レイテンシが安定しない**



OpenvSwitch + DPDK + vhost-user

- **特徴**
  - OpenvSwitch と DPDK (Data Plane Development Kit) を組み合わせたもの
  - NFV で用いられることも多い
  - ポーリングプロセスを user 空間で動作させ、パケットを処理
  - vhost-user を用いることで shared memory を介してパケットをやり取りできる
- **性能**
  - ～20Gbps
- **Pros.**
  - OpenvSwitch をそのまま使うよりも圧倒的に**低レイテンシ**
  - OpenvSwitch の機能をそのまま利用できる
- **Cons.**
  - ポーリングプロセスのために CPU コアを割り当てる必要がある（余分にリソースが必要）

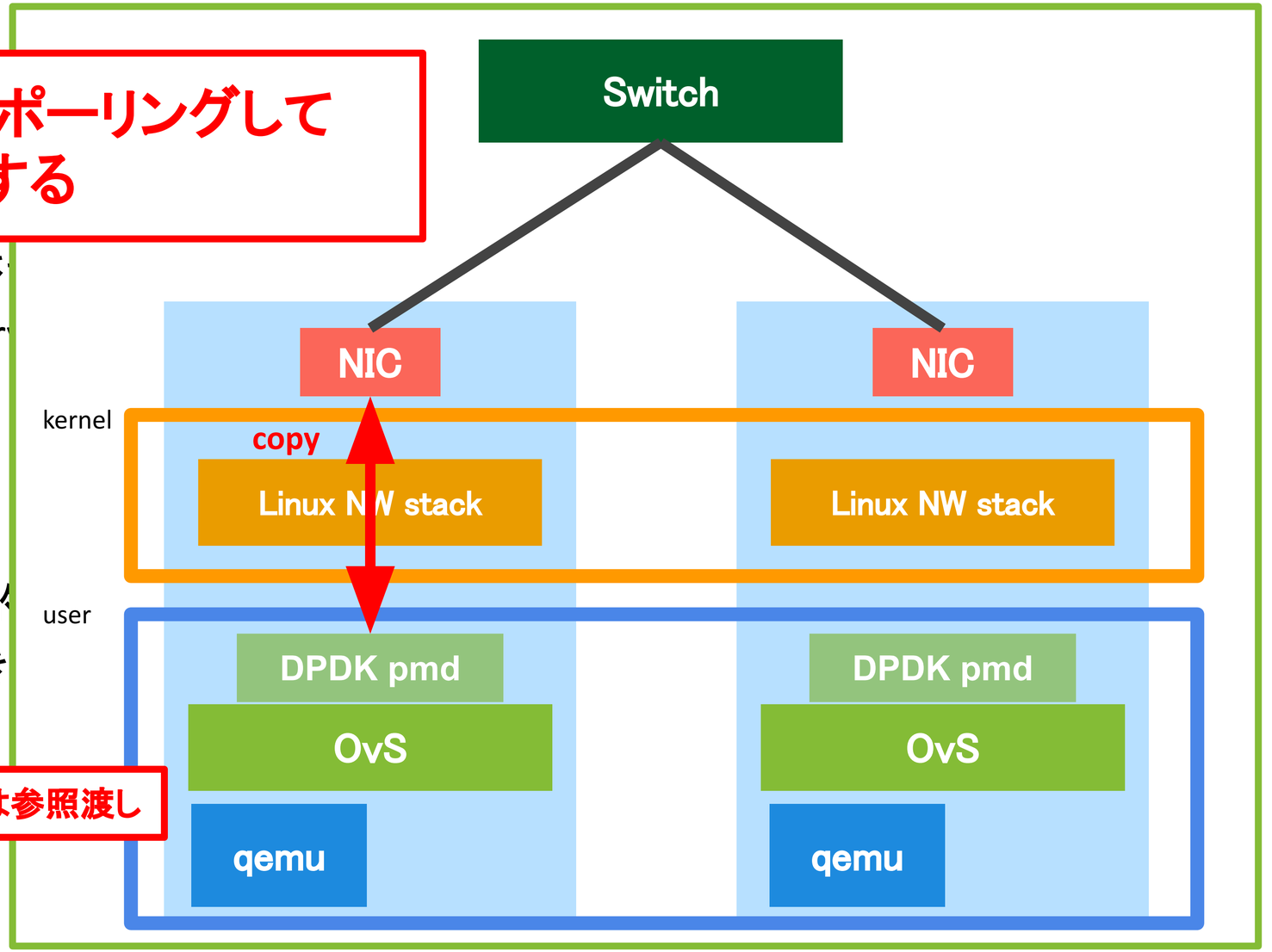
OpenvSwitch + DPDK + vhost-user

user 空間にあるプロセスで NIC をポーリングして kernel 空間をバイパスする

- ポーリングプロセスを user 空間で動作さ
- vhost-user を用いることで shared memor

- **性能**
  - ～20Gbps
- **Pros.**
  - OpenvSwitch をそのまま使うよりも圧倒的
  - OpenvSwitch の機能をそのまま利用でき
- **Cons.**
  - ポーリングプロセスのため

OvS - vNIC 間は参照渡し

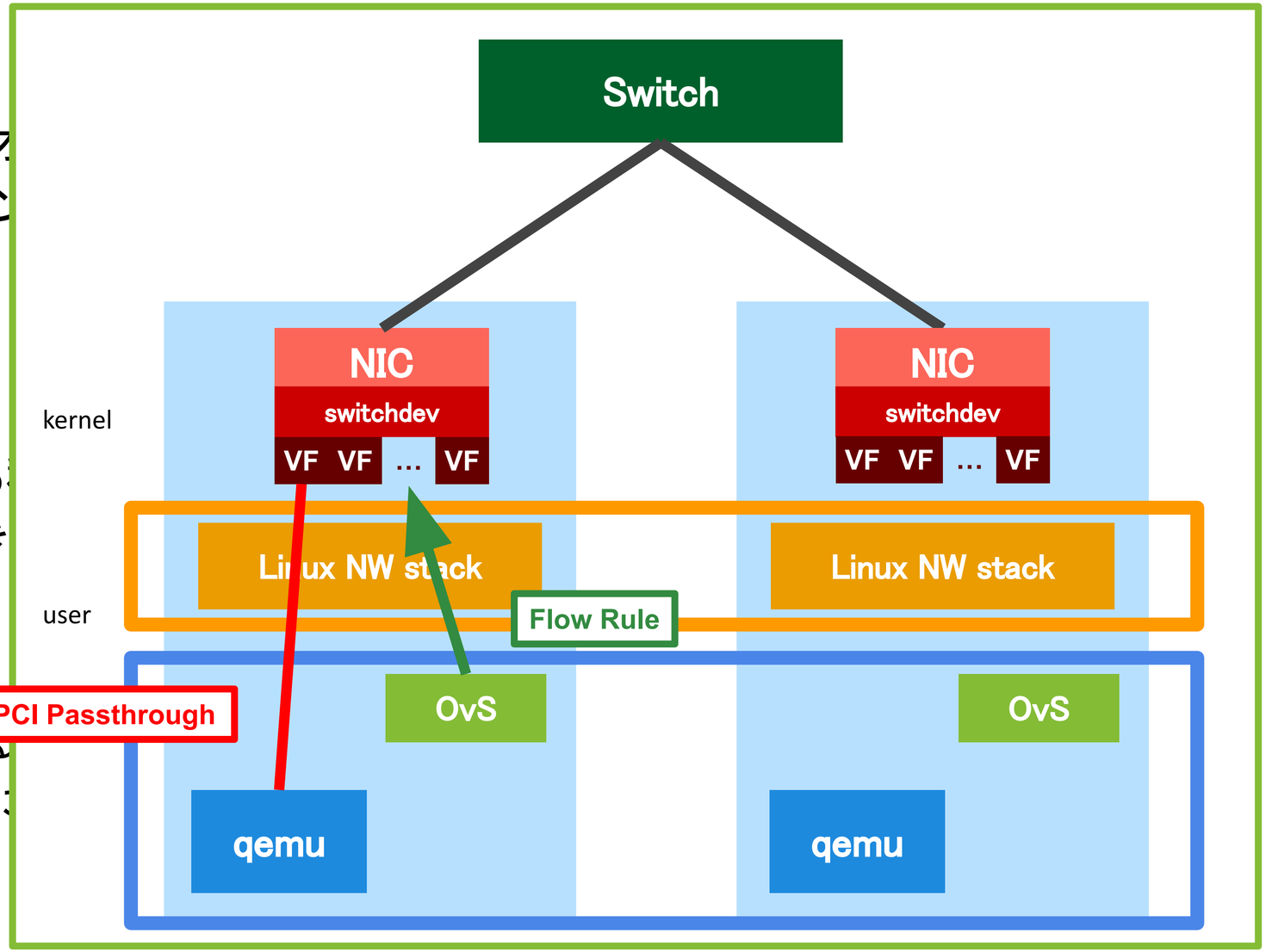


OpenvSwitch HW Offload

- **特徴**
  - OpenvSwitch を NIC 等のハードウェアでオフロードする
  - OpenvSwitch のプロセスは OpenFlow コントローラとしての役割にほぼ専念する
- **性能**
  - ～200Gbps (wirelate)
- **Pros.**
  - 100Gbps x2 でワイヤーレートを目指せる程度に高速
  - OpenvSwitch の機能をそのまま利用できる
- **Cons.**
  - ライブマイグレーションができない
  - 使い方によってはベンダーロックインする
  - VF (SR-IOV で分割したデバイス) の利用にドライバが必要（古い OS だと疎通しない）

OpenvSwitch HW Offload

- **特徴**
  - OpenvSwitch を NIC 等のハードウェアで
  - OpenvSwitch のプロセスは OpenFlow コン
- **性能**
  - ～200Gbps (wirelate)
- **Pros.**
  - 100Gbps x2 でワイヤーレートを目指せる
  - OpenvSwitch の機能をそのまま利用でき
- **Cons.**
  - ライブマイグレーションができない
  - 使い方によってはベンダーロックインする
  - VF (SR-IOV で分割したデバイス) の利用に



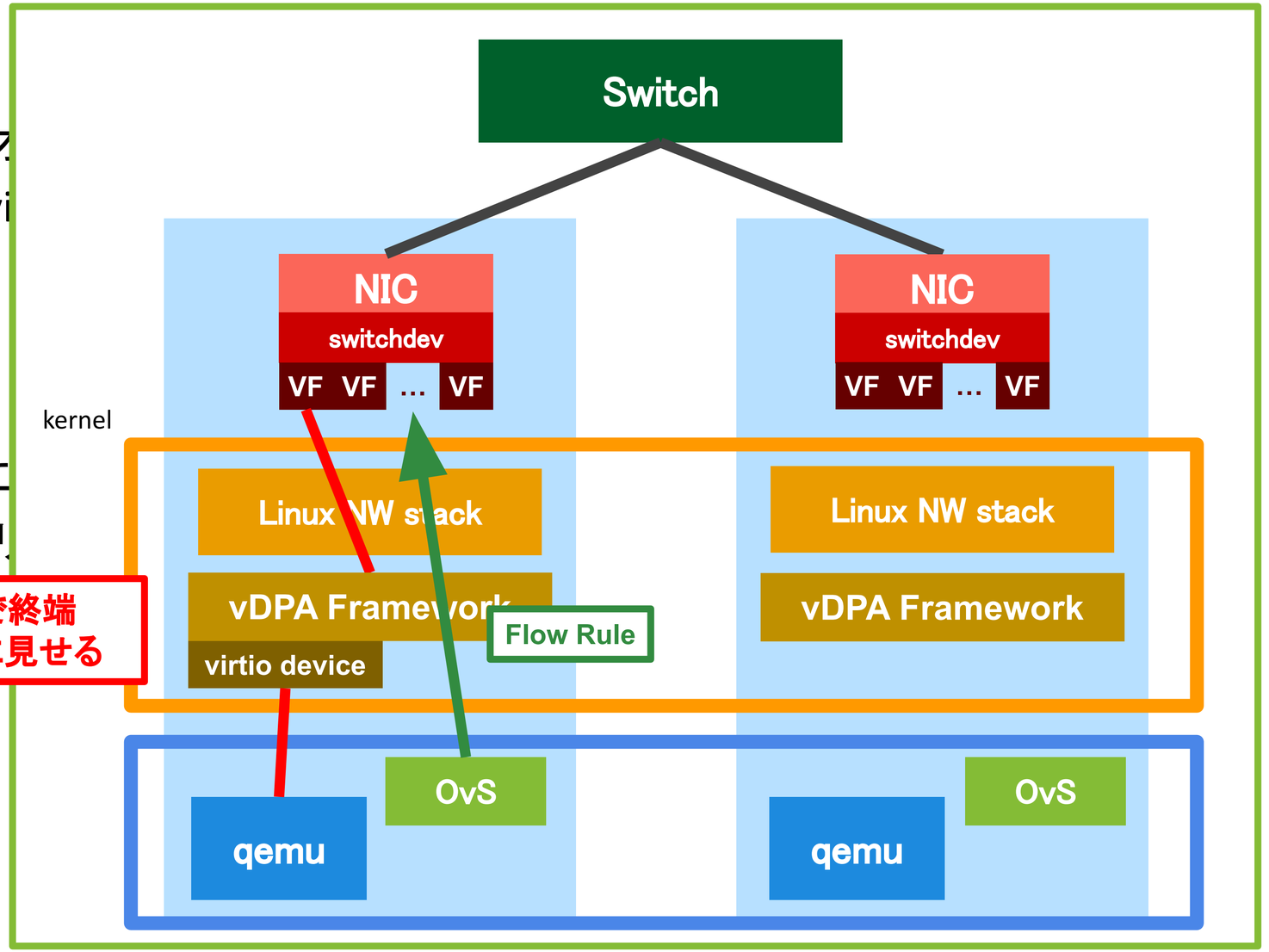
OpenvSwitch + vDPA Kernel Framework

- **特徴**
  - OpenvSwitch を NIC 等のハードウェアでオフロードする
  - VM へ VF を渡さず host kernel で終端、virtio デバイスを渡す
- **性能**
  - ～200Gbps？
- **Pros.**
  - OpenvSwitch HW Offload をより汎用的に利用できる
  - virtio driver さえあれば HW Offload のメリットを享受できる
- **Cons.**
  - ドキュメントが揃っていない

OpenvSwitch + vDPA Kernel Framework

- **特徴**
  - OpenvSwitch を NIC 等のハードウェアで
  - VM へ VF を渡さず host kernel で終端、vi
- **性能**
  - ～200Gbps？
- **Pros.**
  - OpenvSwitch HW Offload をより汎用的に
  - virtio driver さえあれば HW Offload のメ
- **Cons.**
  - ドキュメントが揃っていない

VF を vDPA Framework で終端
virtio device として qemu に見せる



構成まとめ

	linuxbridge	OpenvSwitch	OvS + dpdk	OvS HW offload	OvS + vDPA Framework
Latency	Mid	High	Low	Low	Low
Speed	20Gbps	20Gbps	20Gbps	200Gbps (wirelate)	200Gbps (wirelate)
Live Migration	Yes	Yes	Yes	No	Yes
virtio support	Yes	Yes	Yes	No	Yes
vendor lock	No	No	No	Yes	No
Manageability	High	Mid	Mid	Low	?
CPU usage	Mid	Mid	High (using dedicated core)	Low	Low

構成まとめ

	linuxbridge	OpenvSwitch	OvS + dpdk	OvS HW offload	OvS + vDPA Framework
Latency	Mid	High	Low	Low	Low
Speed	20Gbps	20Gbps	20Gbps	200Gbps (wirelate)	200Gbps (wirelate)
Live Migration	Yes	Yes	Yes	No	Yes
virtio support	Yes	Yes	Yes	No	Yes
vendor lock	No	No	No	Yes	No
Manageability	High	Mid	Mid	Low	?
CPU usage	Mid	Mid	High (using dedicated core)	Low	Low

OpenvSwitch + vDPA Kernel Framework を選択したい！

2

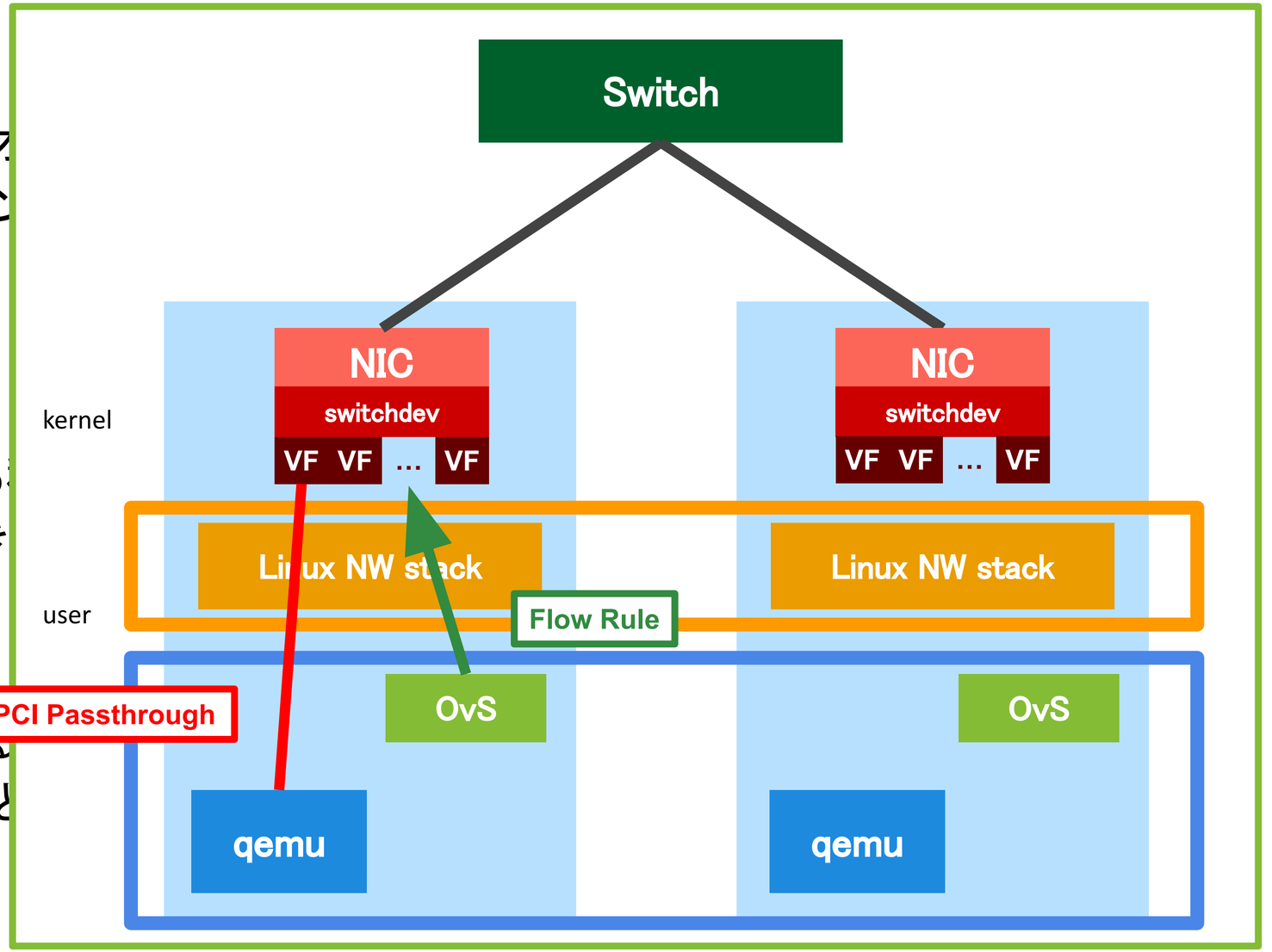
vDPA とは

vDPA とは

- **vDPA とは virtio Data Path Acceleration の略**
  - OpenvSwitch のアクセラレーションとセットで語られることが多い
  - NIC の SR-IOV データプレーンを標準化・抽象化するアプローチ
    - 近いアプローチに **virtio Full HW Offload** がある
      - NIC に virtio エミュレーションを実装
      - （これを vDPA と呼んでる資料もある）
- **vDPA の種類**
  - **vDPA DPDK Framework**
    - DPDK を利用してデータプレーンの抽象化を実現
      - virtio-pmd というプロセスで VF-vhostuser 間をやりとりする
  - **vDPA Kernel Framework**
    - Kernel のフレームワークとして抽象化を実現
  - その他ベンダー独自実装 vDPA

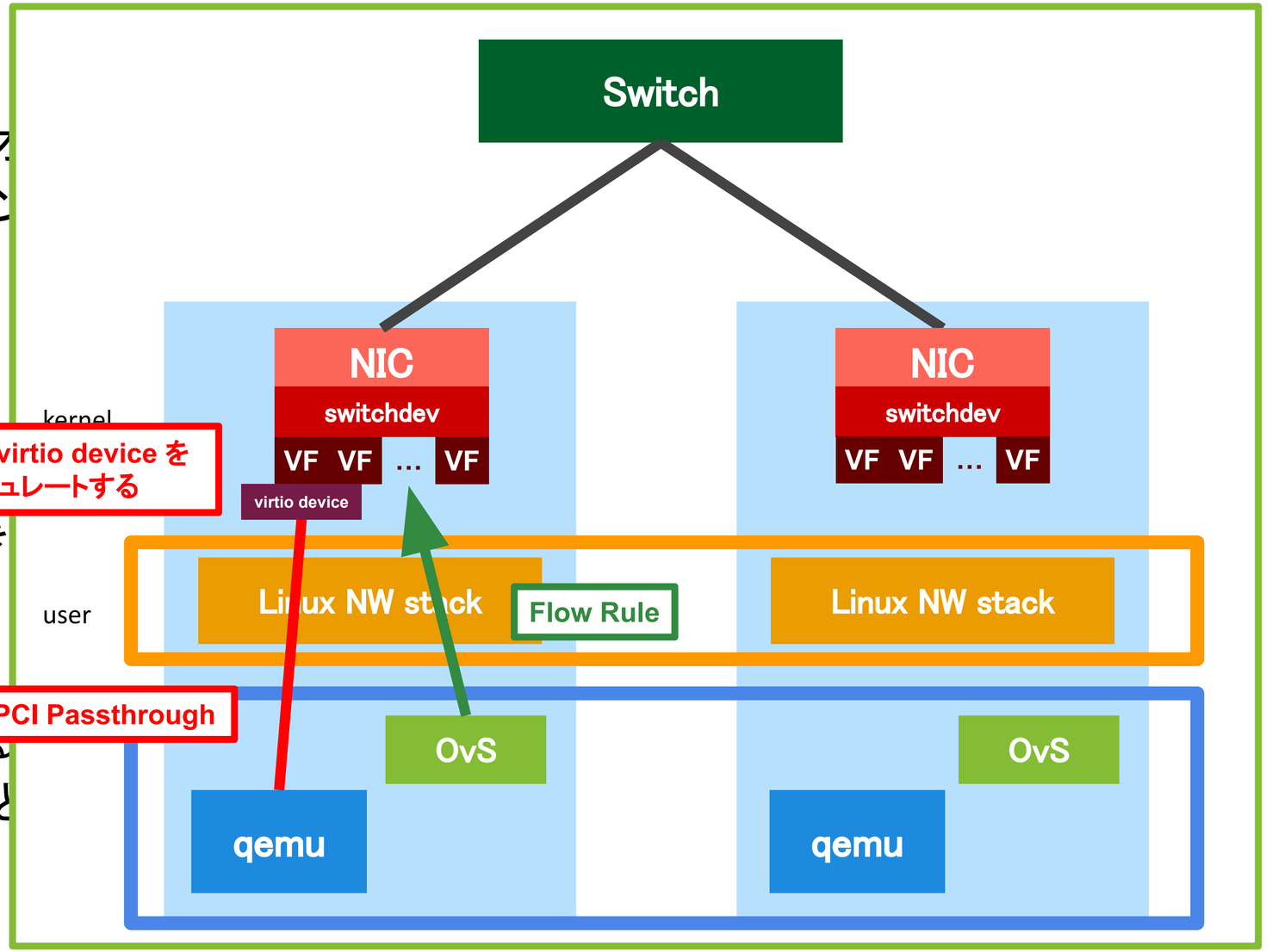
（再掲）OpenvSwitch HW Offload

- **特徴**
  - OpenvSwitch を NIC 等のハードウェアで
  - OpenvSwitch のプロセスは OpenFlow コン
- **性能**
  - ～200Gbps (wirelate)
- **Pros.**
  - 100Gbps x2 でワイヤーレートを目指せる
  - OpenvSwitch の機能をそのまま利用でき
- **Cons.**
  - ライブマイグレーションができない
  - 使い方によってはベンダーロックインする
  - VF の利用にドライバが必要（古い OS だと



OpenvSwitch + virtio Full HW Offload

- **特徴**
  - OpenvSwitch を NIC 等のハードウェアで
  - OpenvSwitch のプロセスは OpenFlow コン
- **性能**
  - ～200Gbps (wirelate)
- **Pros.**
  - 100Gbps x2 でワイヤーレートを目
  - OpenvSwitch の機能をそのまま利用でき
- **Cons.**
  - ライブマイグレーションができない
  - 使い方によってはベンダーロックインする
  - VF の利用にドライバが必要（古い OS だと

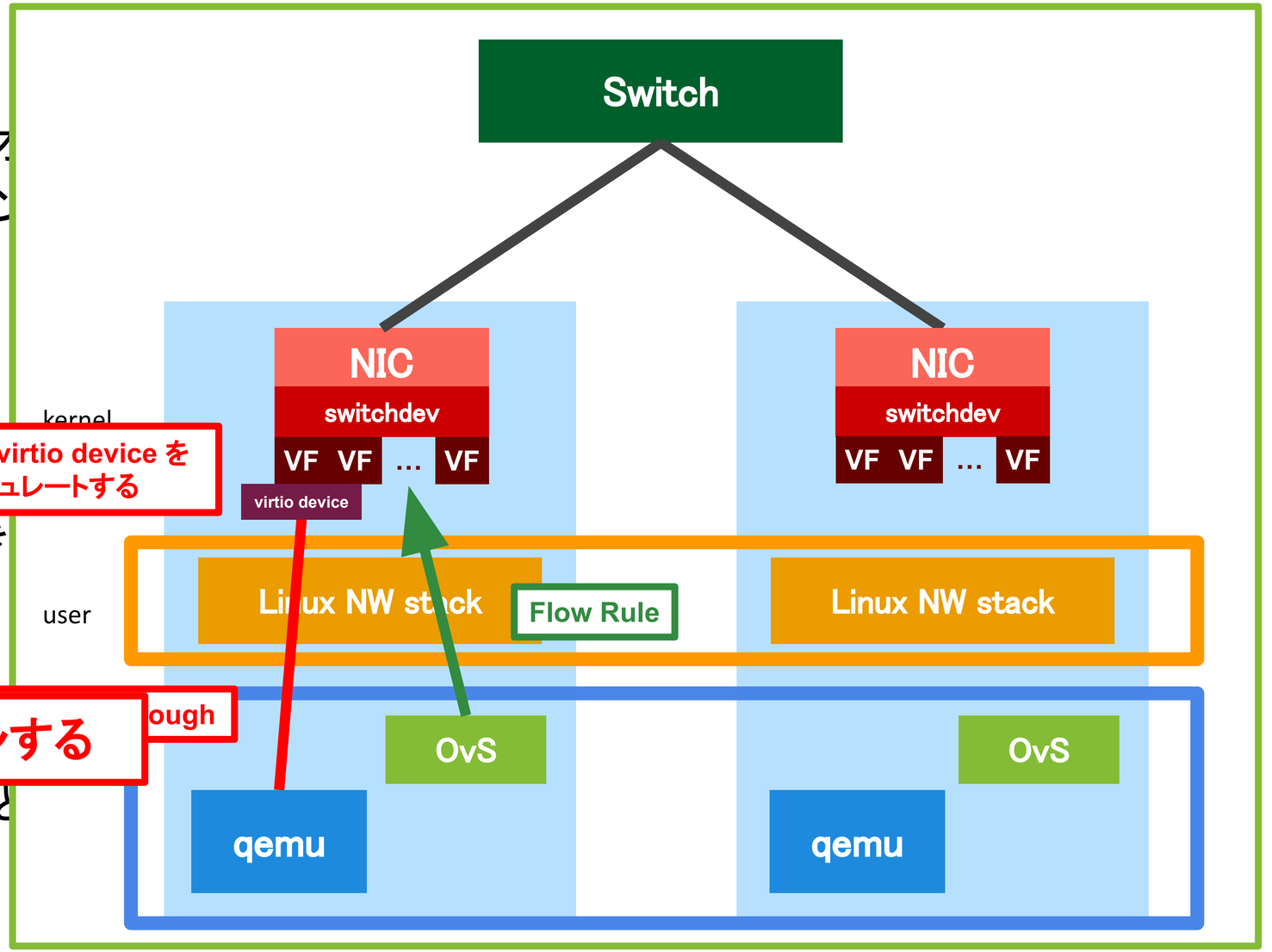


OpenvSwitch + virtio Full HW Offload

- **特徴**
  - OpenvSwitch を NIC 等のハードウェアで
  - OpenvSwitch のプロセスは OpenFlow コン
- **性能**
  - ～200Gbps (wirelate)
- **Pros.**
  - 100Gbps x2 でワイヤーレートを目
  - OpenvSwitch の機能をそのまま利用でき
- **Cons.**
  - ライブマイグレーションができない
  - 使い方によってはパ
  - VF の利用にドライバが必要（古い OS だと

NIC が virtio device をエミュレートする

ベンダーロックインする



OpenvSwitch + virtio Full HW Offload

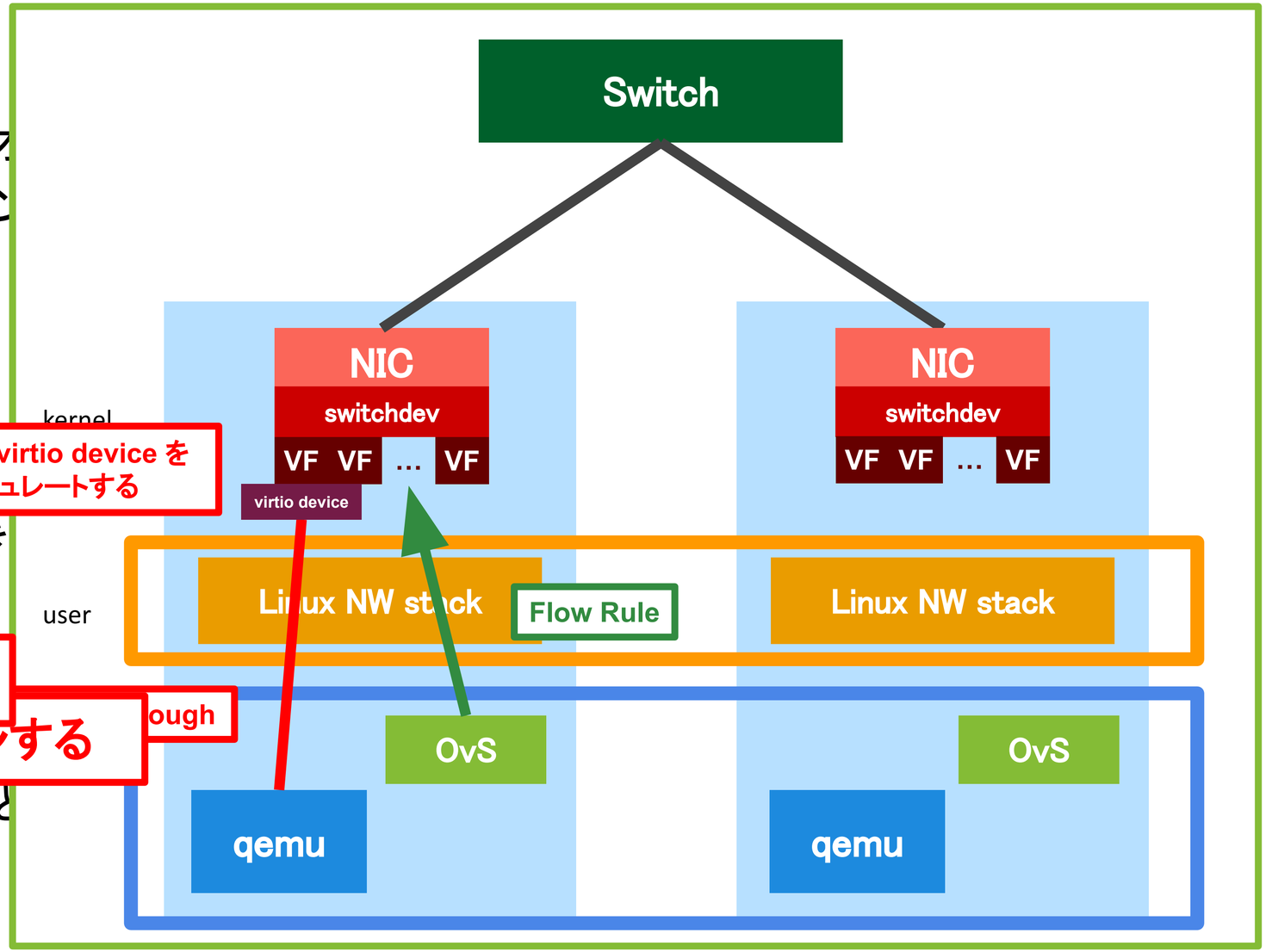
- **特徴**
  - OpenvSwitch を NIC 等のハードウェアで
  - OpenvSwitch のプロセスは OpenFlow コン
- **性能**
  - ～200Gbps (wirelate)
- **Pros.**
  - 100Gbps x2 でワイヤーレートを目
  - OpenvSwitch の機能をそのまま利用でき
- **Cons.**
  -
  - 使い方によっては
  - VF の利用にドライバが必要（古い OS だと

NIC が virtio device をエミュレートする

ライブマイグレーションができない

ough

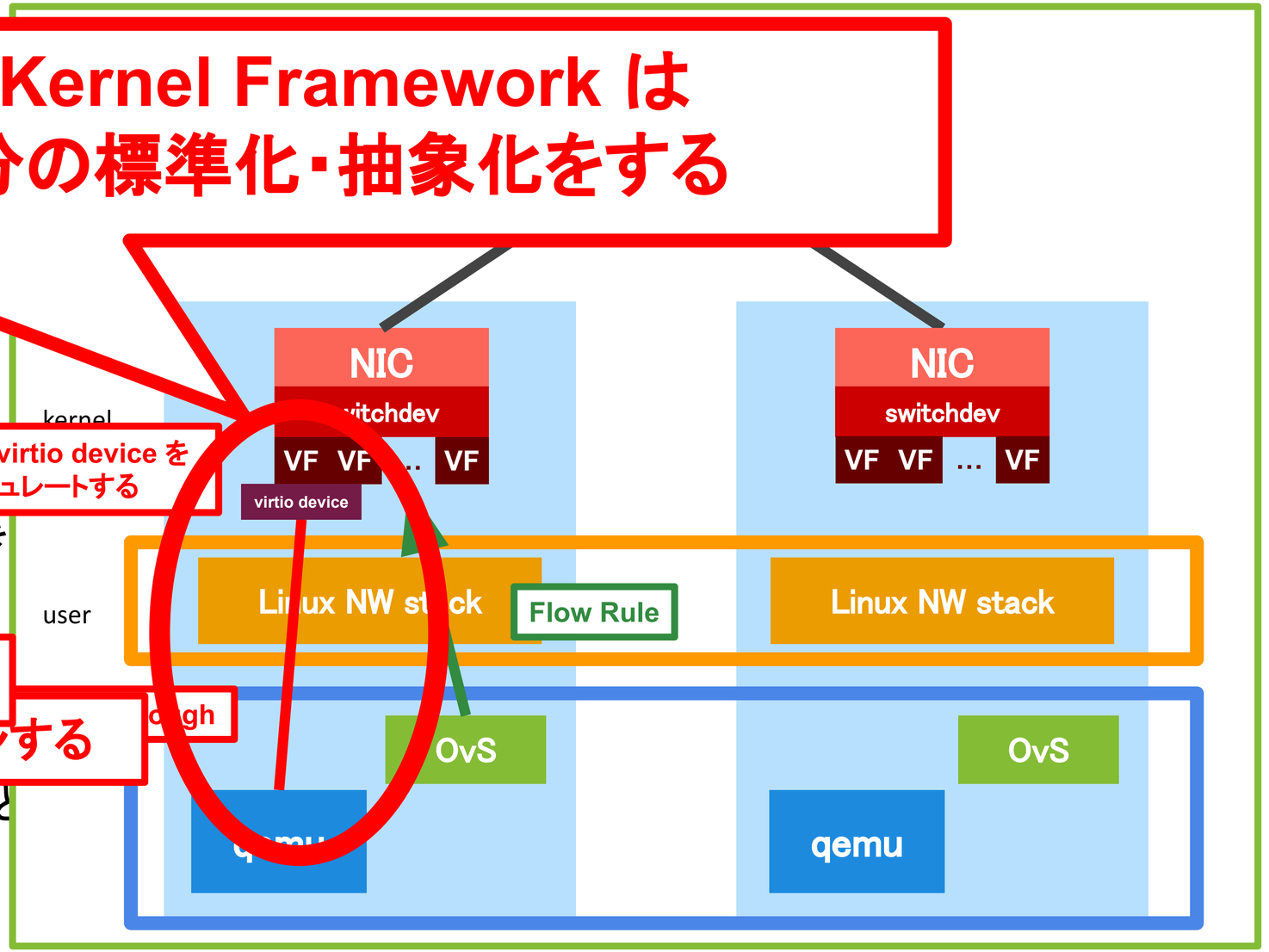
ベンダーロックインする



OpenvSwitch + virtio Full HW Offload

- **特徴**
  - OpenvSwitch を
  - OpenvSwitch の
- **性能**
  - ～200Gbps (wirelate)
- **Pros.**
  - 100Gbps x2 でワイヤーレートを目
  - OpenvSwitch の機能をそのまま利用でき
- **Cons.**
  - ライブマイグレーションができない
  - 使い方によってはベンダーロックインする
  - VF の利用にドライバが必要（古い OS だと

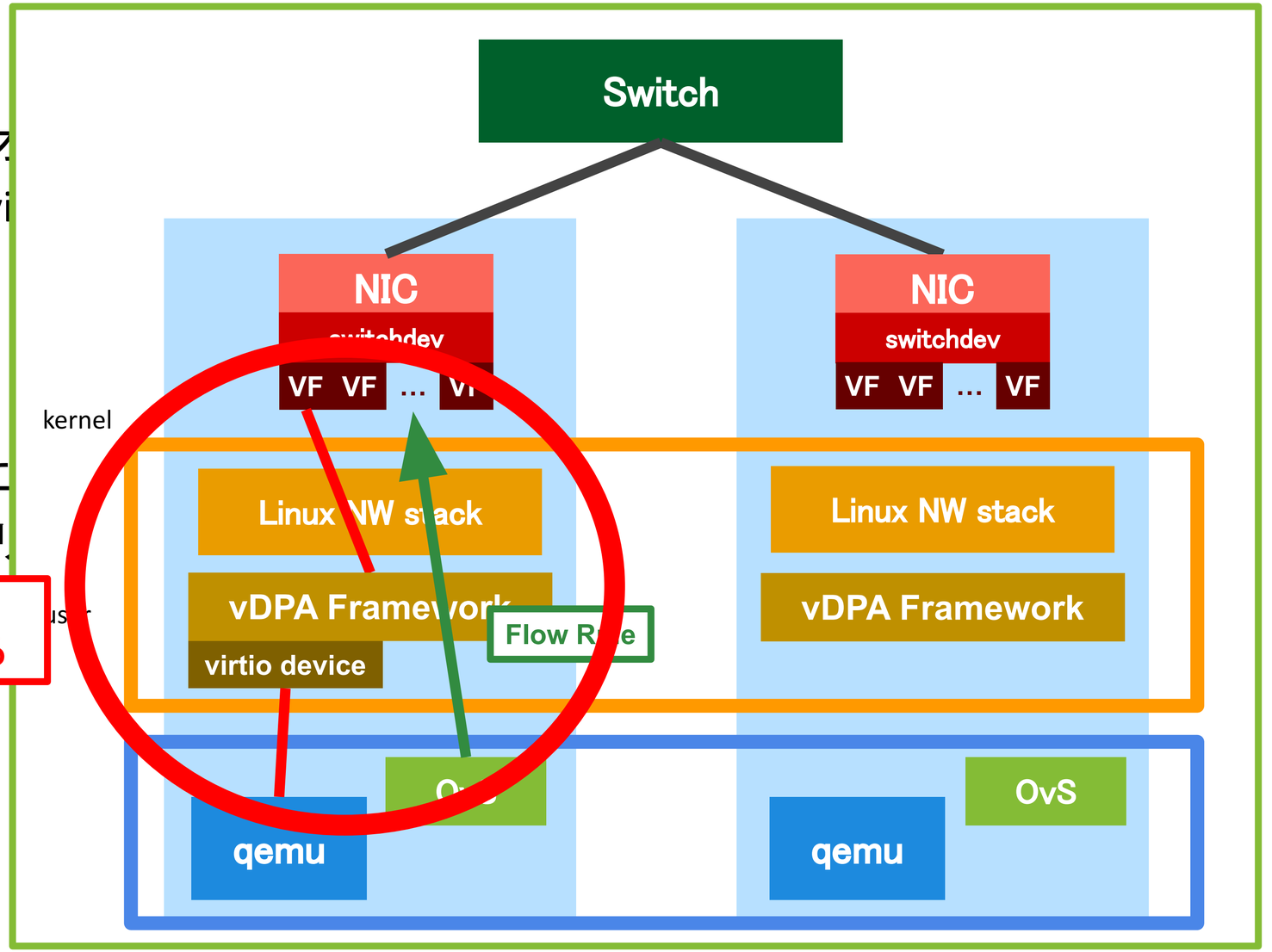
vDPA Kernel Framework は
この部分の標準化・抽象化をする



（再掲）OpenvSwitch + vDPA Kernel Framework

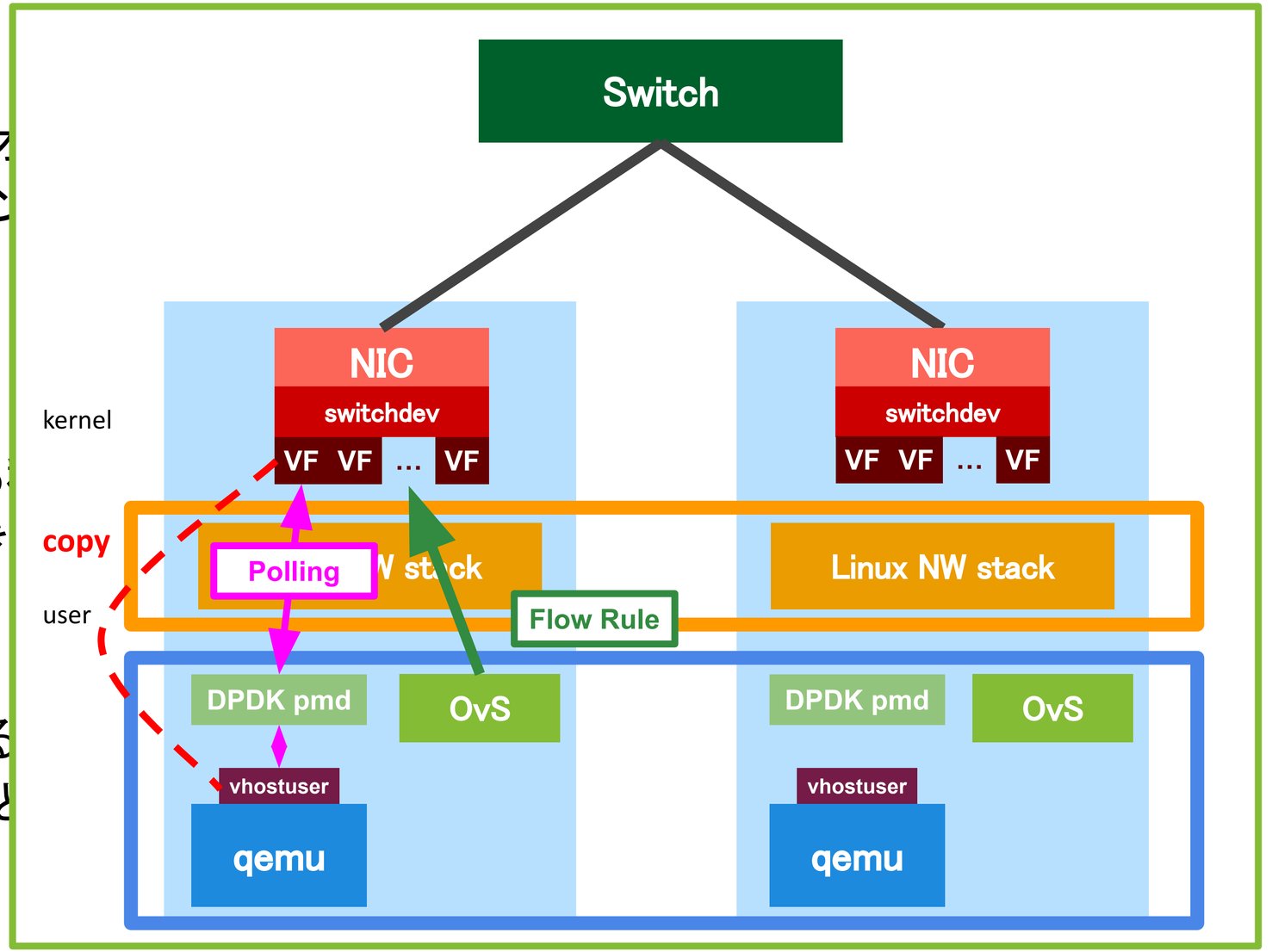
- **特徴**
  - OpenvSwitch を NIC 等のハードウェアで
  - VM へ VF を渡さず host kernel で終端、vi
- **性能**
  - ～200Gbps？
- **Pros.**
  - OpenvSwitch HW Offload をより汎用的に
  - virtio driver さえあれば HW Offload のメ
- **Cons.**
  - ドキュメントが揃っていない

VF を vDPA Framework で終端
virtio device として qemu に見せる



OpenvSwitch + vDPA DPDK Framework

- **特徴**
  - OpenvSwitch を NIC 等のハードウェアで
  - OpenvSwitch のプロセスは OpenFlow コン
- **性能**
  - ～200Gbps (wirelate)
- **Pros.**
  - 100Gbps x2 でワイヤーレートを目指せる
  - OpenvSwitch の機能をそのまま利用でき
- **Cons.**
  - ライブマイグレーションができない
  - 使い方によってはベンダーロックインする
  - VF の利用にドライバが必要（古い OS だと



3

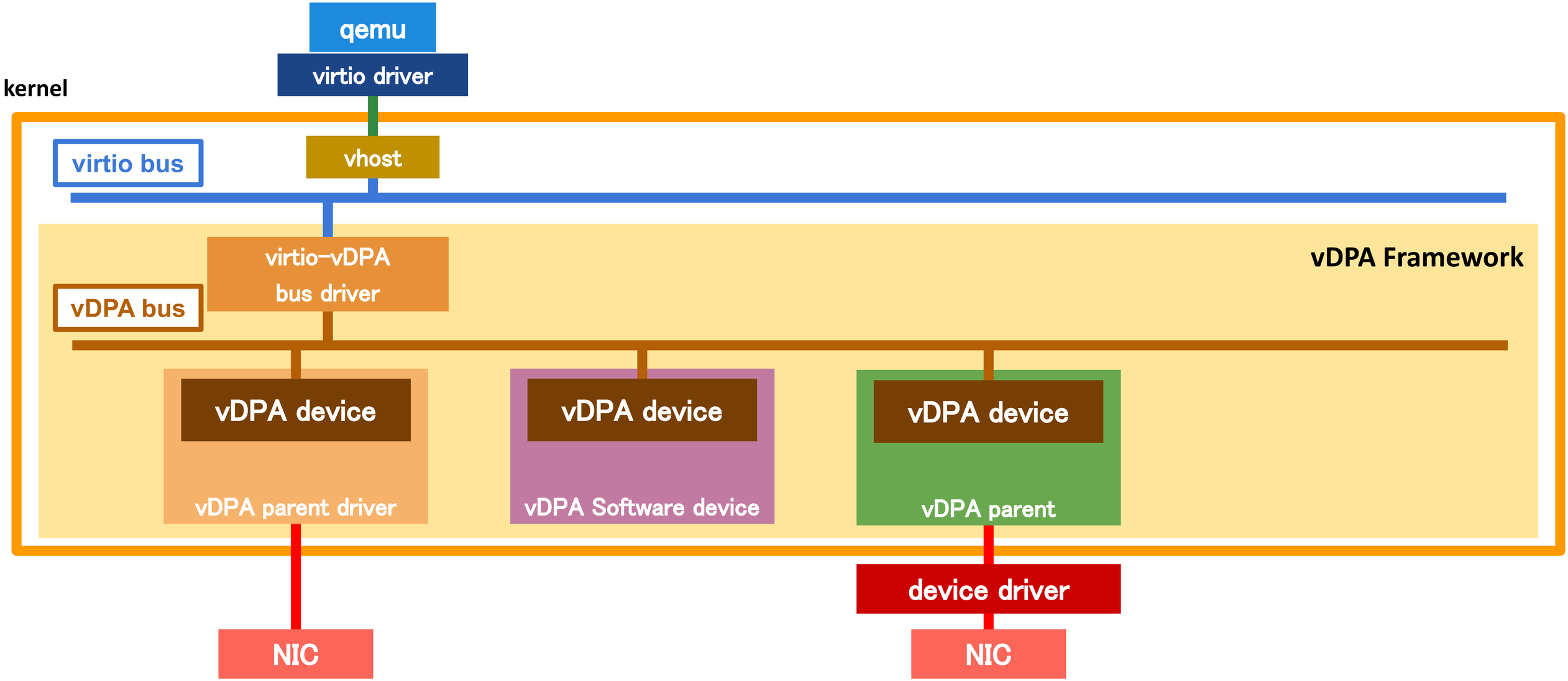
vDPA Kernel Framework

vDPA Kernel Framework

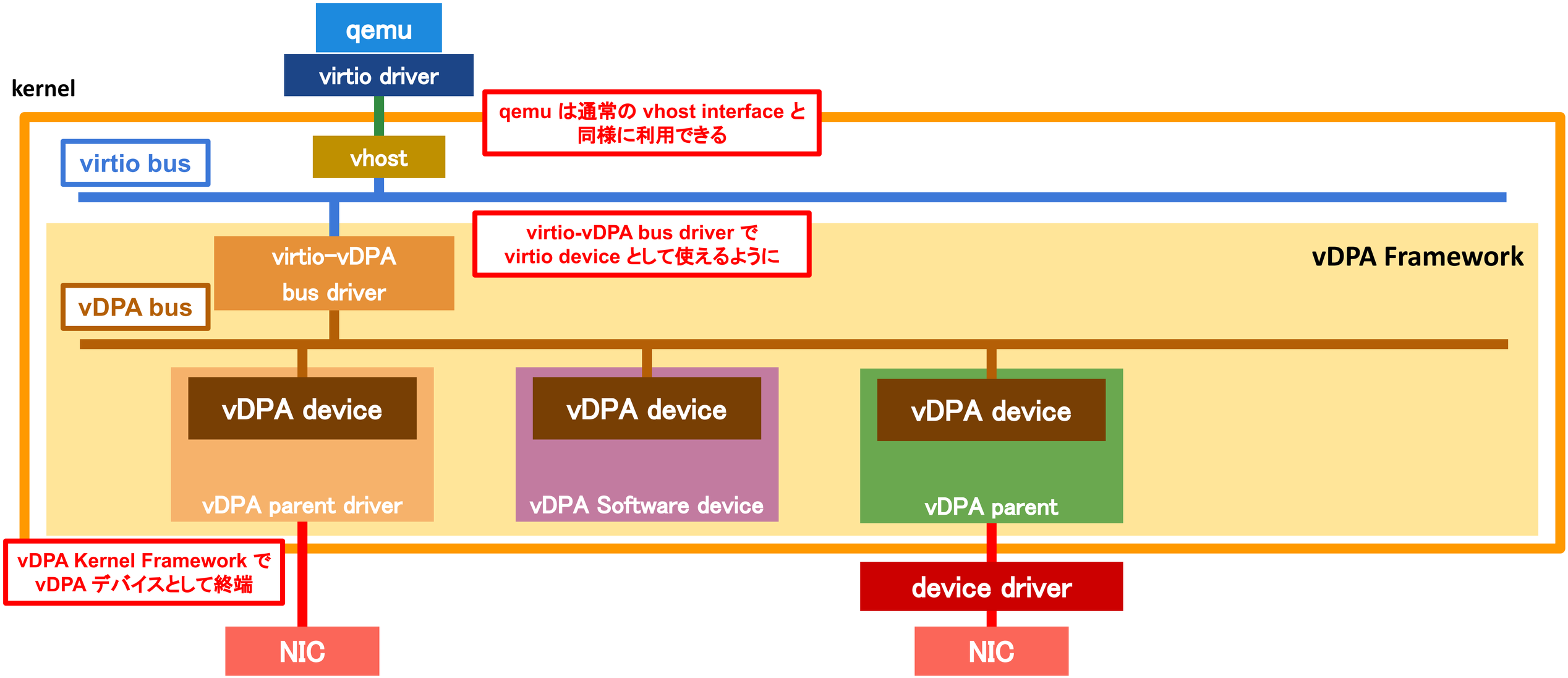
- **vDPA の標準化・抽象化のための Kernel Framework**
  - Linux Kernel 5.7 でマージされた
  - これまでベンダー毎に実装していた vDPA を抽象化する
    - ベンダーの負担が減る／仕様が標準化されることでユーザにもメリットがある
      - ベンダー・製品の選択肢が増える／ベンダーロックインしにくい

- **virtio driver だけでどんなデバイスでも使えるように**
  - ベアメタルやコンテナでも virtio driver のみでデバイスを使えるようになる
  - 今回の例はネットワークだが、ストレージでもこの仕組みを使うことができる

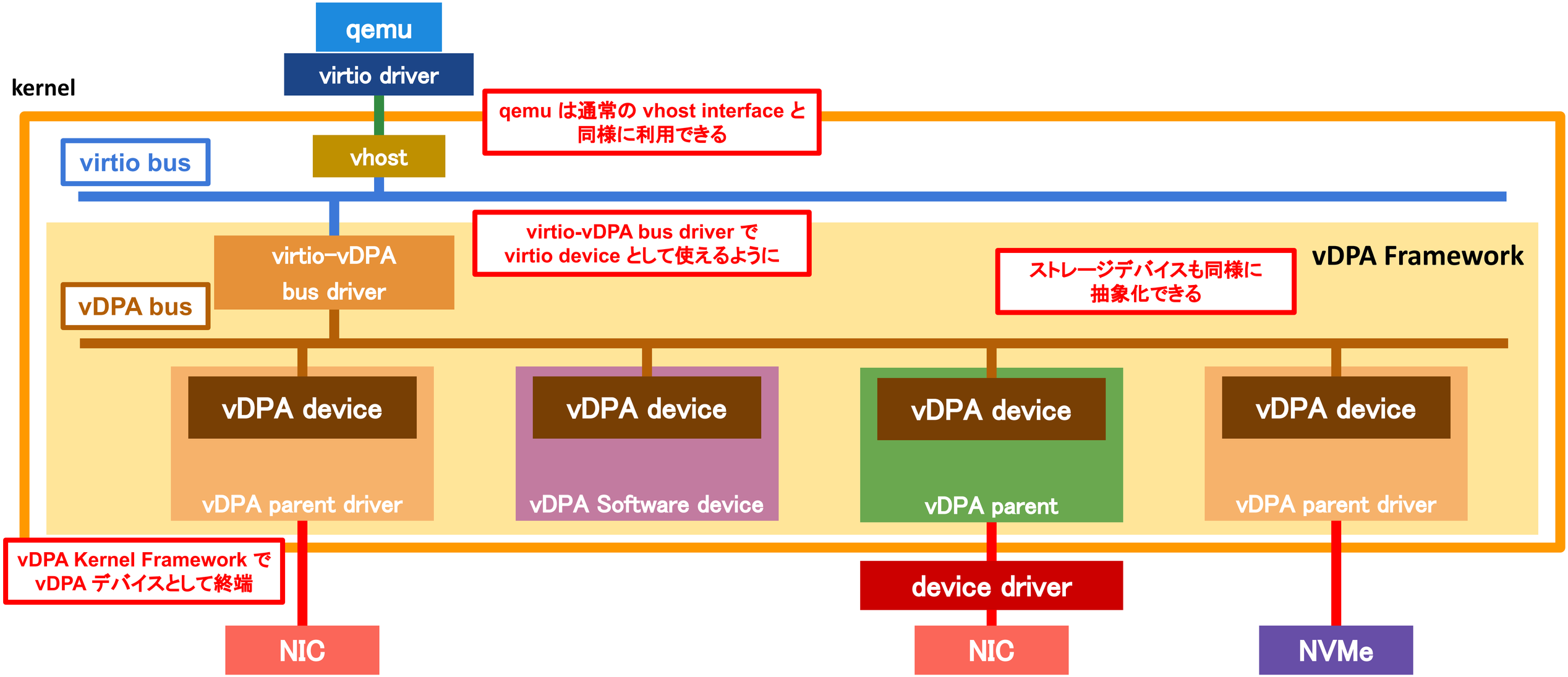
Overview


qemu
virtio driver
kernel
virtio bus
vhost
virtio-vDPA
bus driver
vDPA bus
vDPA Framework
vDPA device
vDPA device
vDPA device
vDPA parent driver
vDPA Software device
vDPA parent
device driver
NIC
NIC

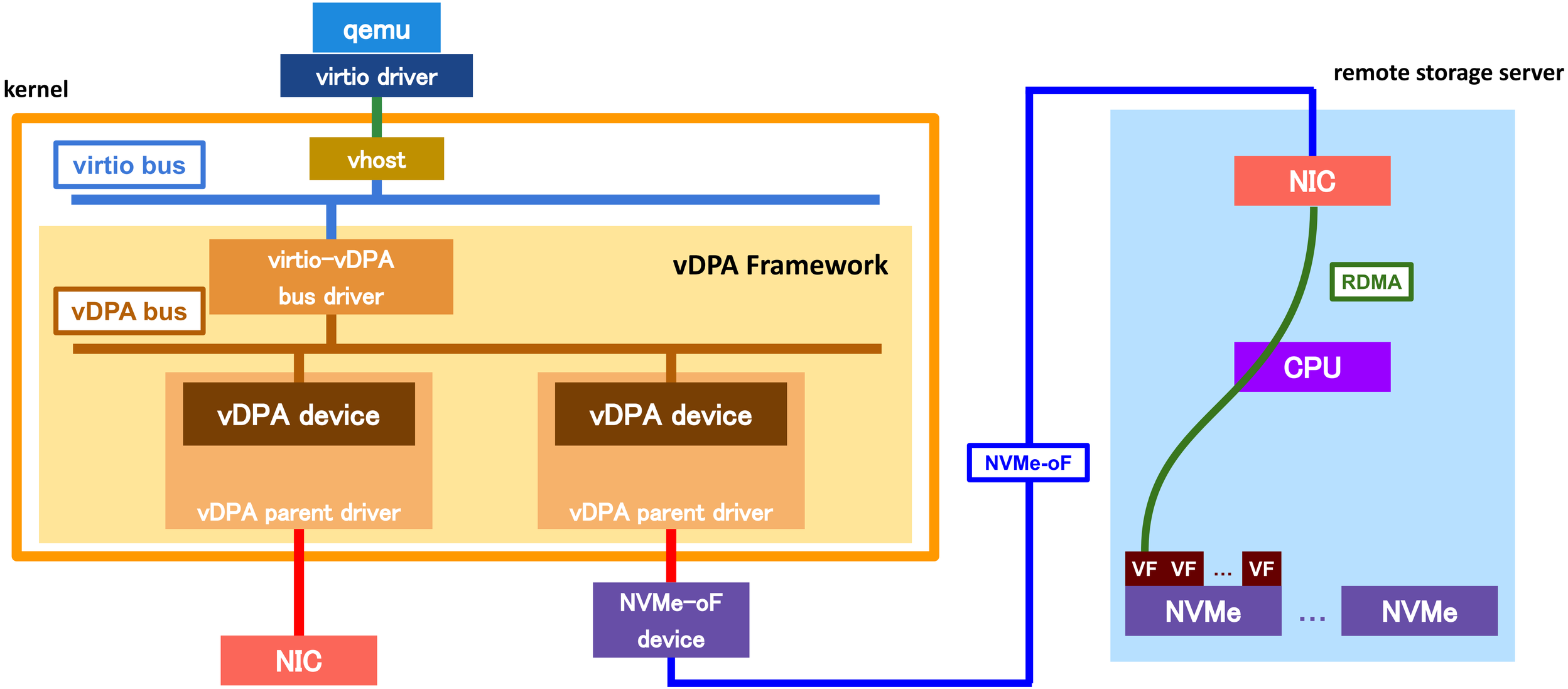
Overview


qemu
virtio driver
kernel
qemu は通常の vhost interface と
同様に利用できる
virtio bus
vhost
virtio-vDPA bus driver で
virtio device として使えるように
virtio–vDPA
bus driver
vDPA Framework
vDPA bus
vDPA device
vDPA device
vDPA device
vDPA parent driver
vDPA Software device
vDPA parent
vDPA Kernel Framework で
vDPA デバイスとして終端
device driver
NIC
NIC

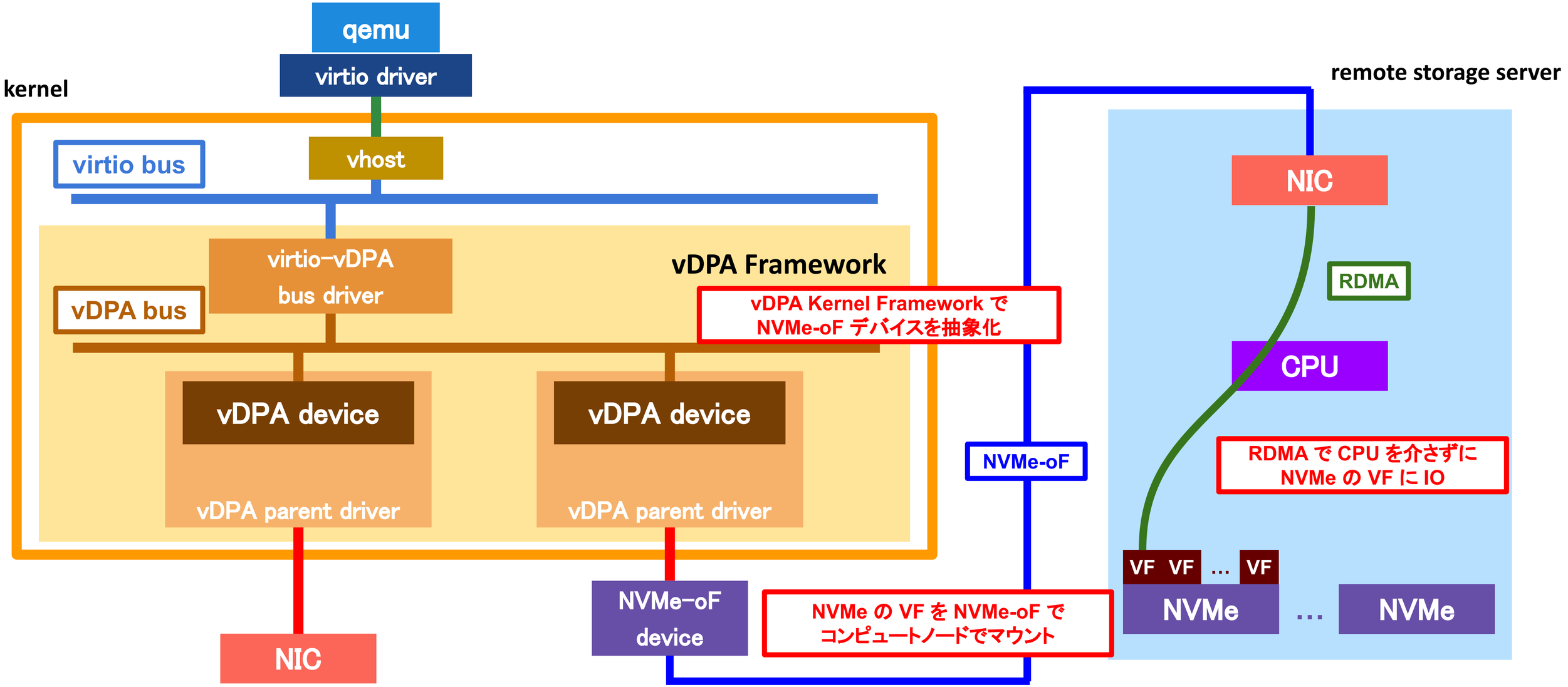
Overview


qemu
virtio driver
kernel
qemu は通常の vhost interface と
同様に利用できる
virtio bus
vhost
virtio-vDPA bus driver で
virtio device として使えるように
virtio–vDPA
bus driver
vDPA Framework
ストレージデバイスも同様に
抽象化できる
vDPA bus
vDPA device
vDPA device
vDPA device
vDPA device
vDPA parent driver
vDPA Software device
vDPA parent
vDPA parent driver
vDPA Kernel Framework で
vDPA デバイスとして終端
device driver
NIC
NIC
NVMe

vDPA with NVMe-oF



vDPA with NVMe-oF



4

まとめ

CyberAgent

まとめ

- **HV のネットワーキングには様々なアプローチがある**
  - linuxbridge
  - OpenvSwitch
  - OpenvSwitch + X
- **vDPA Kernel Framework は将来のスタンダードになるか**
  - ワイヤーレート
  - virtio driver さえあれば使える
  - ベンダーロックなし
  - ライブマイグレーション可能
- **vDPA Kernel Framework を利用した IaaS/コンテナ基盤に向けて検証を進めていく**
  - 200Gbps が出せる最強最速 VM 基盤
  - vDPA + NVMe-oF の最強最速ストレージ基盤

Reference

- **Introduction to vDPA kernel framework**
  https://www.redhat.com/ja/blog/introduction-vdpa-kernel-framework
- **Achieving network wirespeed in an open standard manner: introducing vDPA**
  https://www.redhat.com/en/blog/achieving-network-wirespeed-open-standard-manner-introducing-vdpa
- **How deep does the vDPA rabbit hole go?**
  https://www.redhat.com/en/blog/how-deep-does-vdpa-rabbit-hole-go
- **vDPA support in Linux kernel (KVM Forum 2020) - Jason Wang, Red Hat**
  https://www.youtube.com/watch?v=f7IbuKCCKgs

Recruit

- **24卒新卒採用**

  https://www.cyberagent.co.jp/careers/special/students/tech/

- **中途採用**

  https://hrmos.co/pages/cyberagent-group/jobs?category=16140868252919234 56

EOP