Mintオペレーティングシステムにおける仮想ネットワークインタフェースの改善

平成31年2月15日

岡山大学 工学部 情報系学科
吉田 修太郎

背景と目的

Mint:

1台の計算機上で
複数のOS(OSノード)を走行

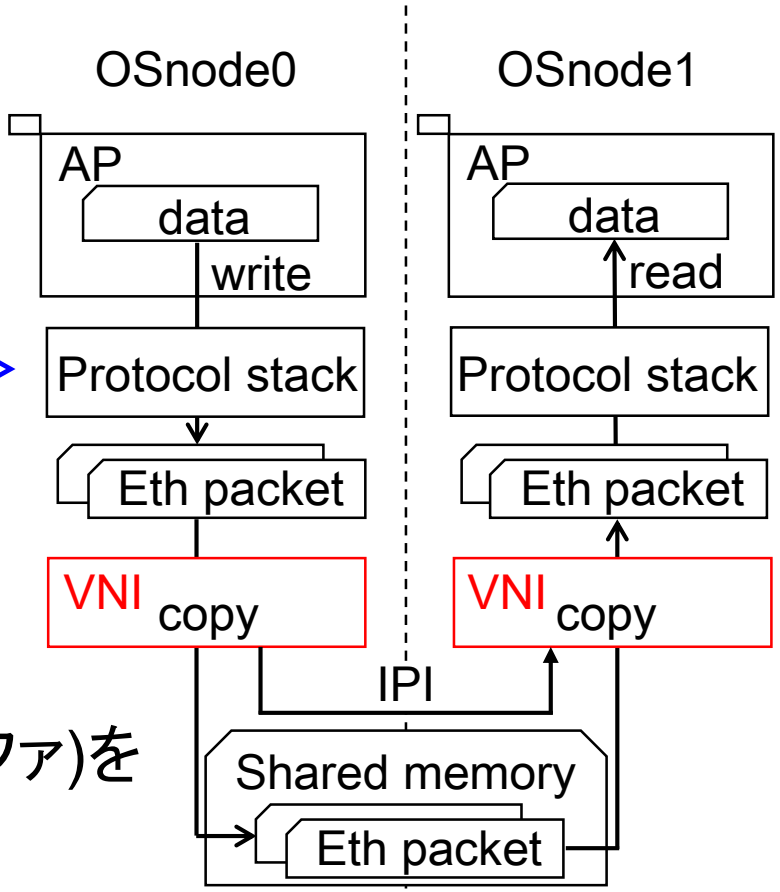
＜仮想ネットワークインタフェース (VNI)＞

(1) OSノード間でEthernet互換の
通信を実現

(2) IPIでハードウェア割込みを代替

(3) 共有メモリの特定領域(送受信バッファ)を
介してパケットを送受信

(4) 通信性能について課題有



既存のVNIにおける課題の解消

既存のVNIにおける課題

(課題1) TCPプロトコルを用いた通信におけるスループットの向上

TCPプロトコルを用いた通信のスループットは約128Mbps

(課題2) OSノード間における排他制御の検討

(A) 共有メモリの使用はOSノード間で排他制御を要する

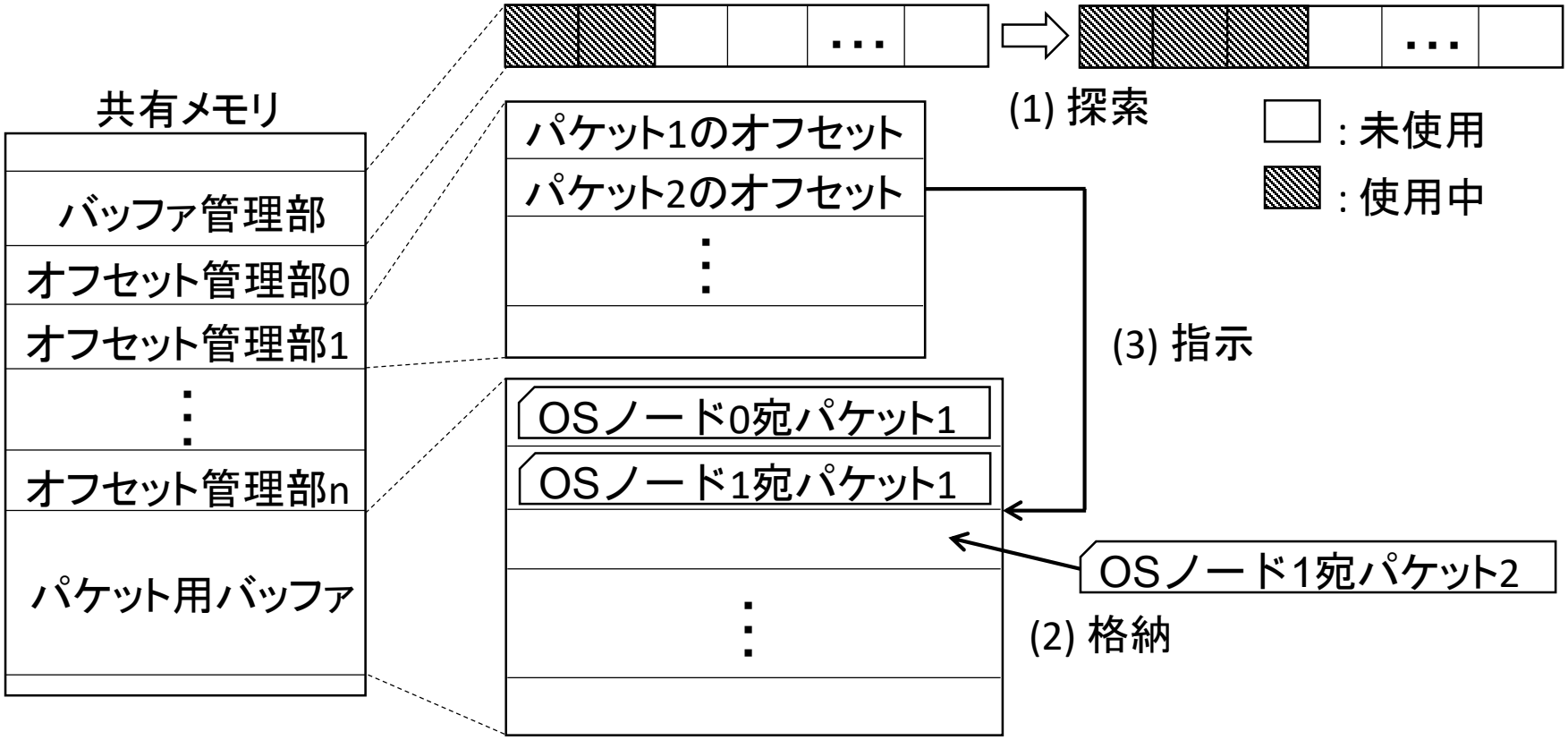
(B) OSノード間における排他制御は未実装であり，
共有メモリへのアクセスを時分割で行わない限り
通信できない

⇒ 通信性能の向上は望めない

⇓

送受信バッファの構成を見直し，共有メモリの使用について
OSノード間における排他制御の必要性を検討

既存の送受信バッファにおける構成と処理流れ

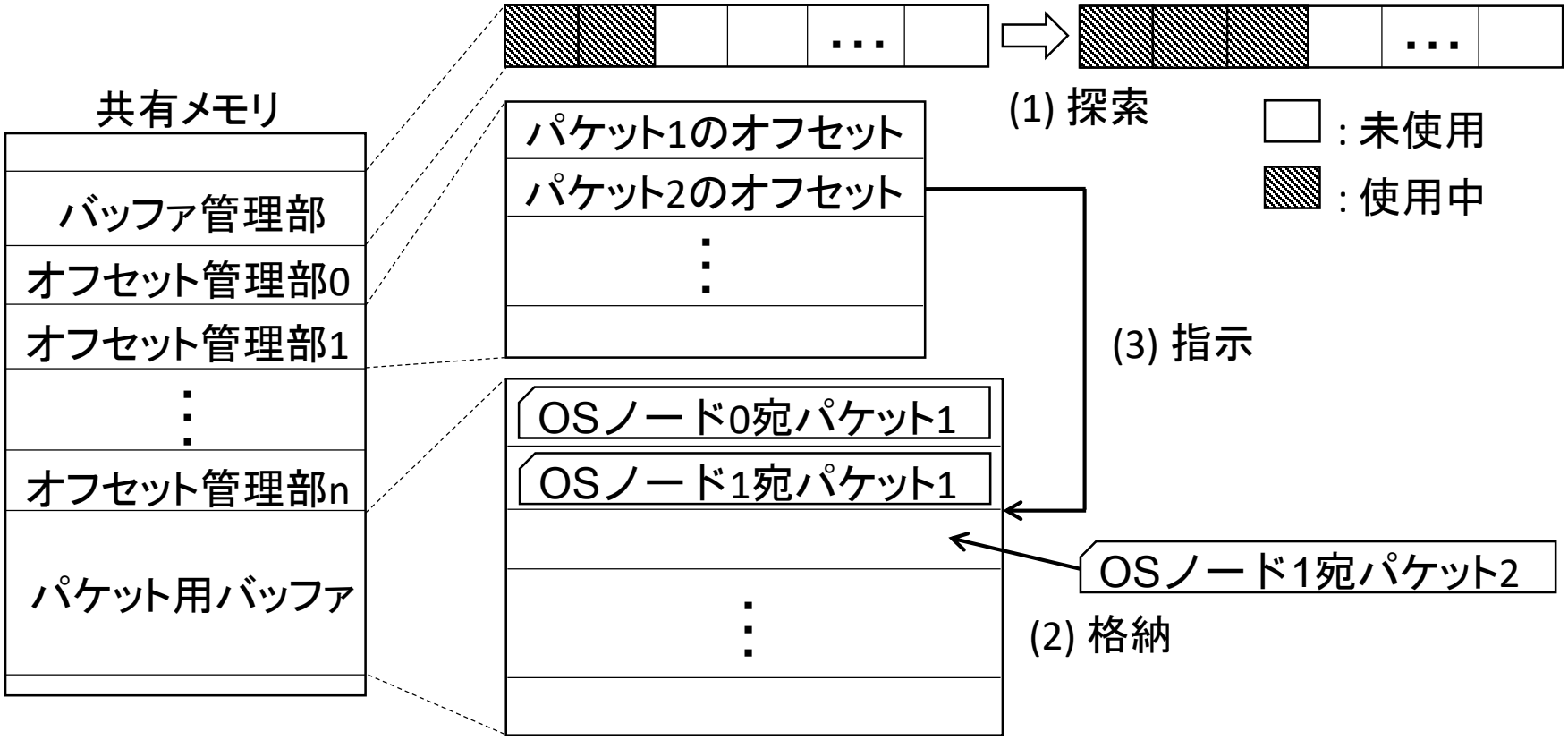


パケット用バッファ: パケットを格納

オフセット管理部: パケットの格納位置を管理

バッファ管理部: パケット用バッファの各領域が使用中か否か管理

既存の送受信バッファにおける構成と処理流れ

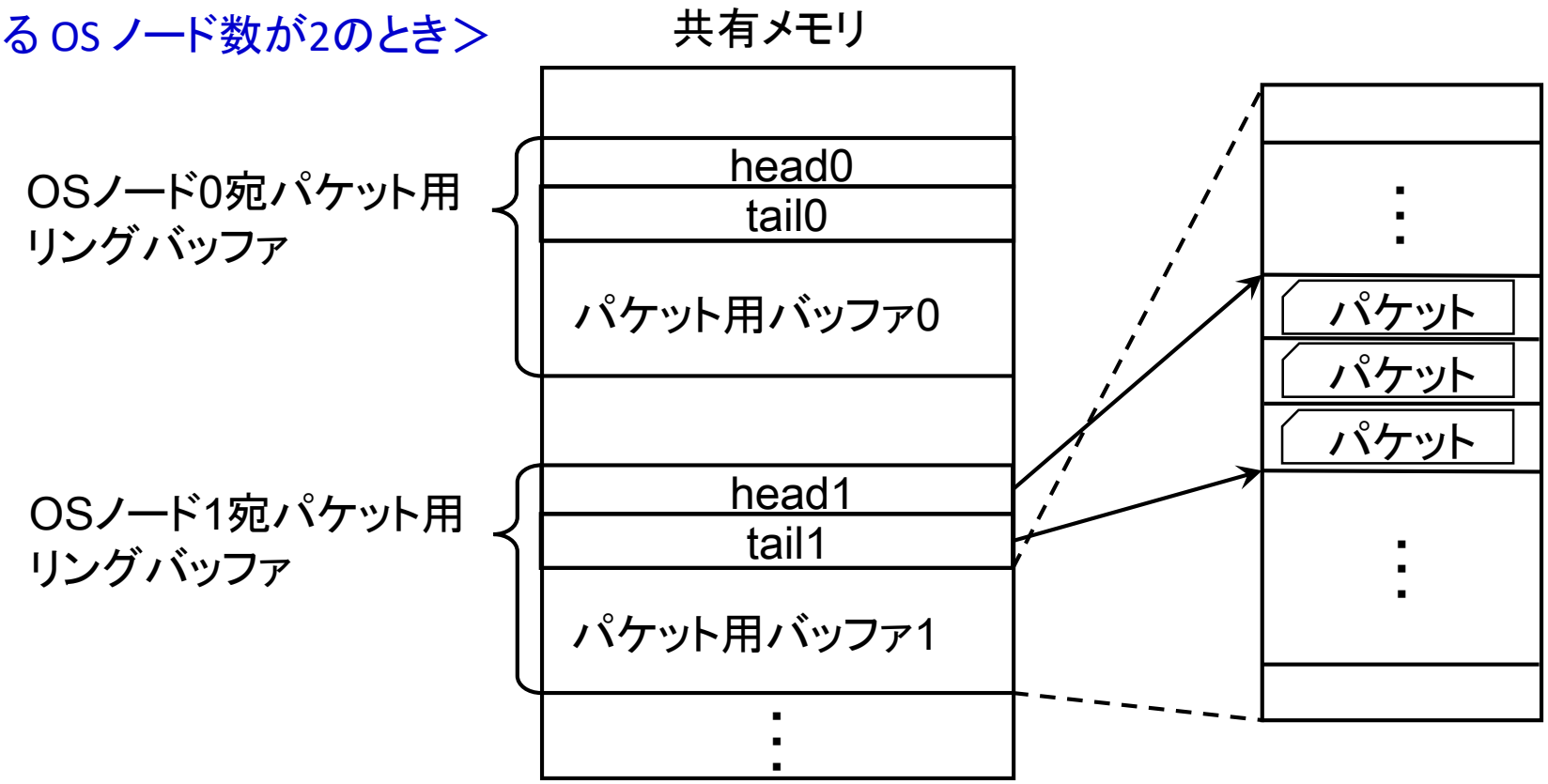


バッファ管理部，オフセット管理部，パケット用バッファのいずれの領域においても，複数のOSノードが同時に書き込む可能性有

OSノード間の競合を生じる可能性有

OS ノード間で競合を生じない送受信バッファ構成

＜通信する OS ノード数が2のとき＞



(1) パケットを格納する領域を宛先ごとに分割する
(2) 送受信バッファにおける1つの分割した領域に対する書込みは，
それぞれ1つのOSノードのみが行う

➡ 送受信バッファの操作はOSノード間で競合しない

スループット計測

＜目的＞

(1) 送受信バッファの改変による, VNIを用いた通信におけるスループットの変化を調査する

＜環境＞

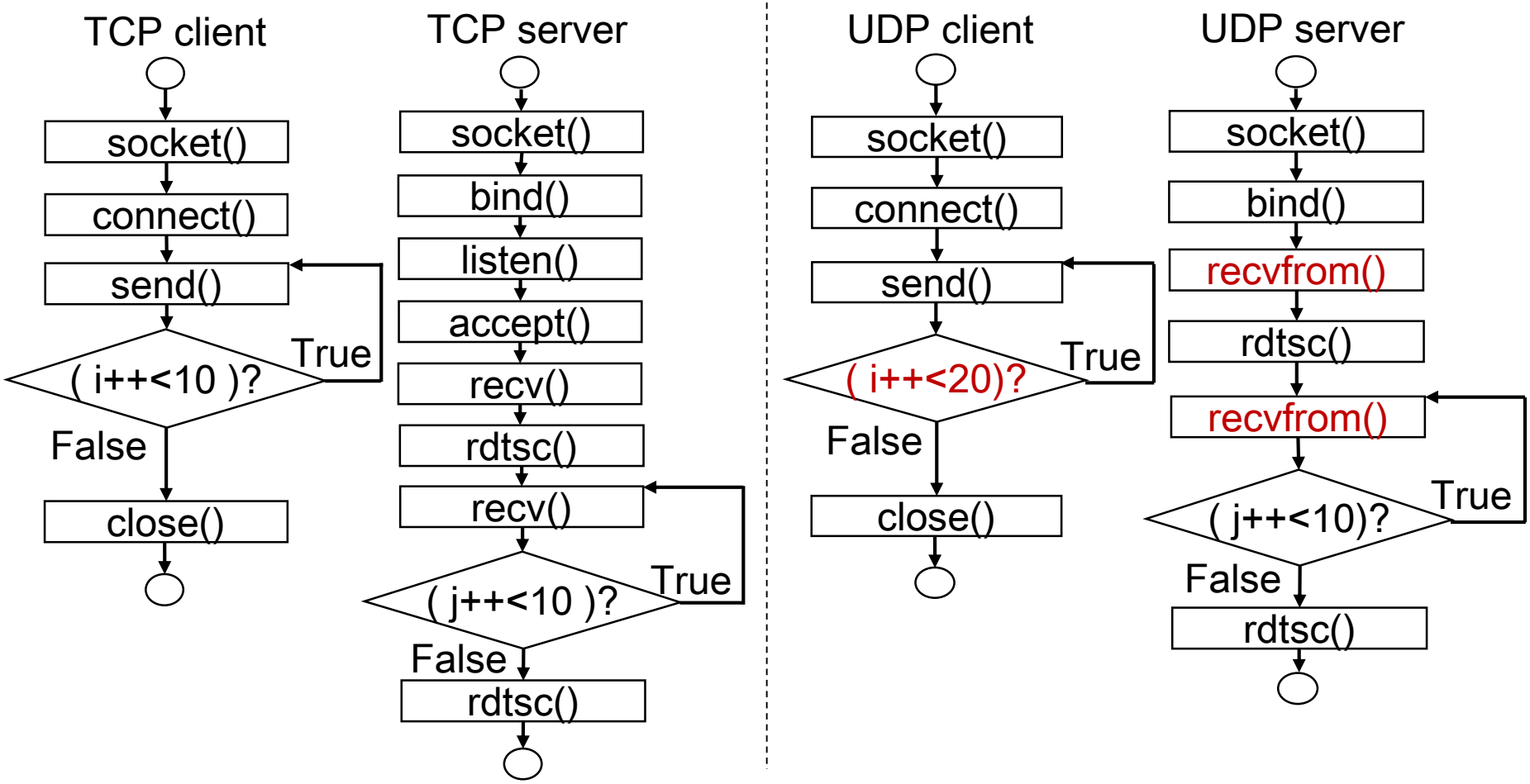
OS	Debian 7.11
カーネル	Mint(Linux kernel v3.15 から改変)
起動するOSノード数	2
各OSノードの持つコア数	1
CPU	Intel Core i7-4770 (3.40GHz)
メモリの容量	16GB
メモリI/Oの帯域幅	25.6GB/sec
共有メモリのサイズ	16MB
送受信バッファのサイズ	392,200B

＜手法＞

(1) 通信 プロトコルとしてUDPプロトコルを用いた場合とTCPプロトコルを用いた場合の2つの場合について通信のスループットを計測

(2) 計測にはユーザプログラムを用いる

計測用プログラムの処理流れ



(1) 1 回 15,000Byte の send()/recv() を 10 回繰り返す

(2) recv() したデータの量を，recv() に要した時間で割る

スループットの計測結果

＜結果＞

送受信バッファ	TCP	UDP
改変前	128Mbps	65Gbps
改変後	10Gbps	75Gbps

＜改変前と改変後の比較＞

(1) UDPプロトコルを用いた通信とTCPプロトコルを用いた通信のどちらの場合においてもスループットは向上

(2) UDPプロトコルを用いた通信と比較して, TCPプロトコルを用いた通信の方がスループットの上昇率が大きい

まとめ

＜実績＞

(1) 送受信バッファにおける構成の再検討

(2) 送受信バッファの再実装

(3) VNI を用いた通信のスループット計測

(A) 既存の VNI による通信

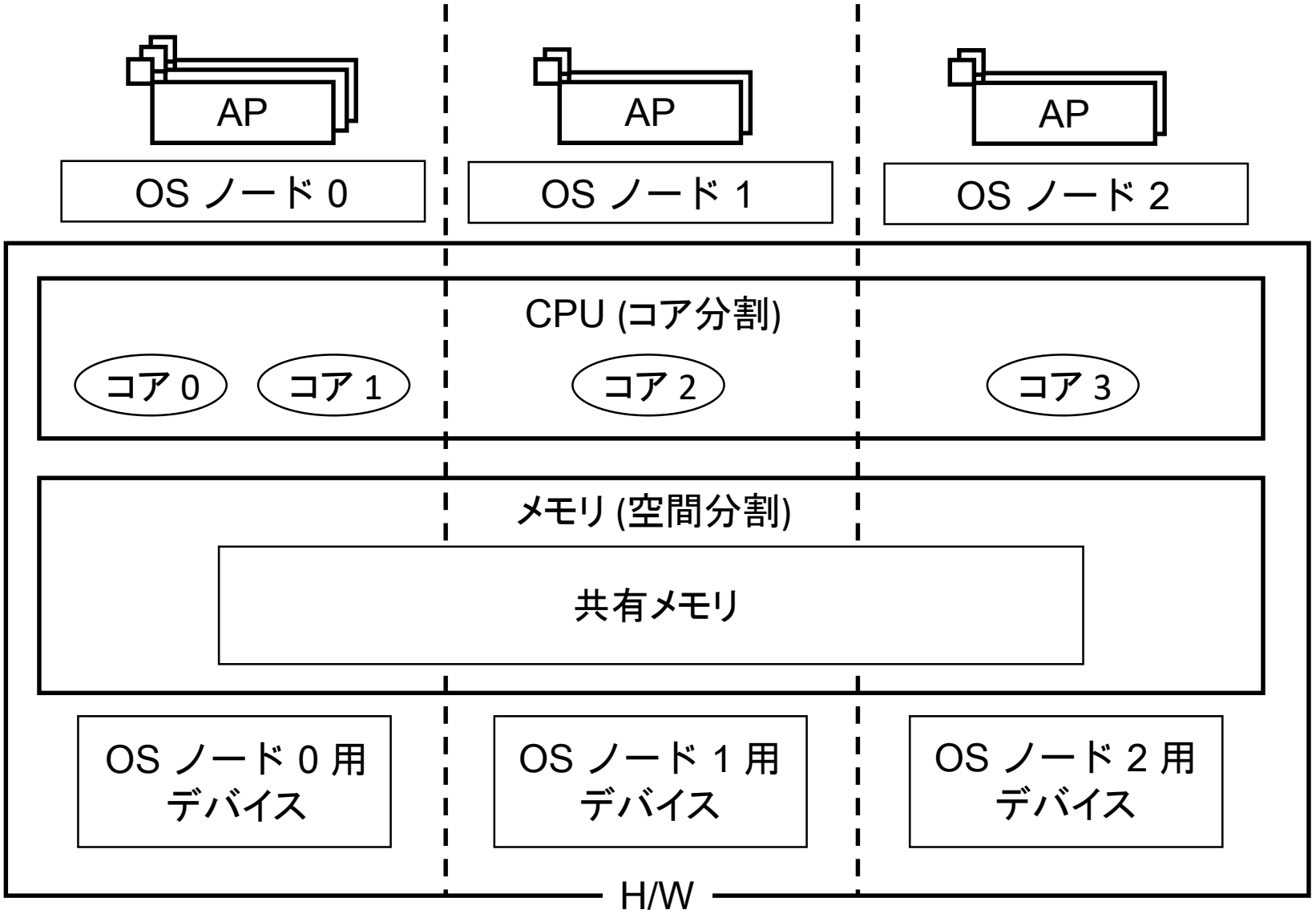
(B) 送受信バッファを再実装したVNIによる通信

＜今後の課題＞

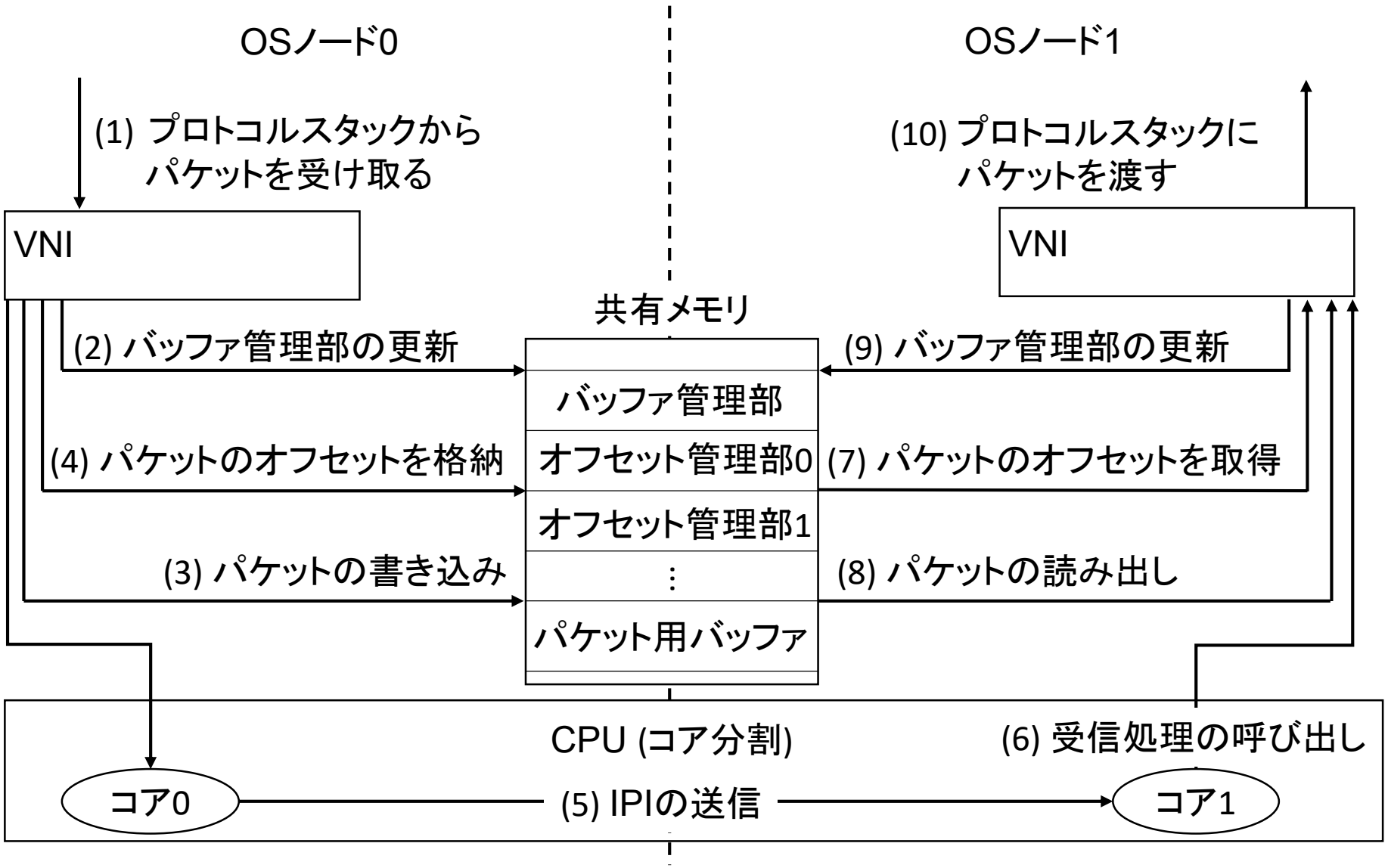
再実装した送受信バッファ構成では，通信するOSノードの増加に伴い，より多くのパケット用バッファを要する

予備スライド

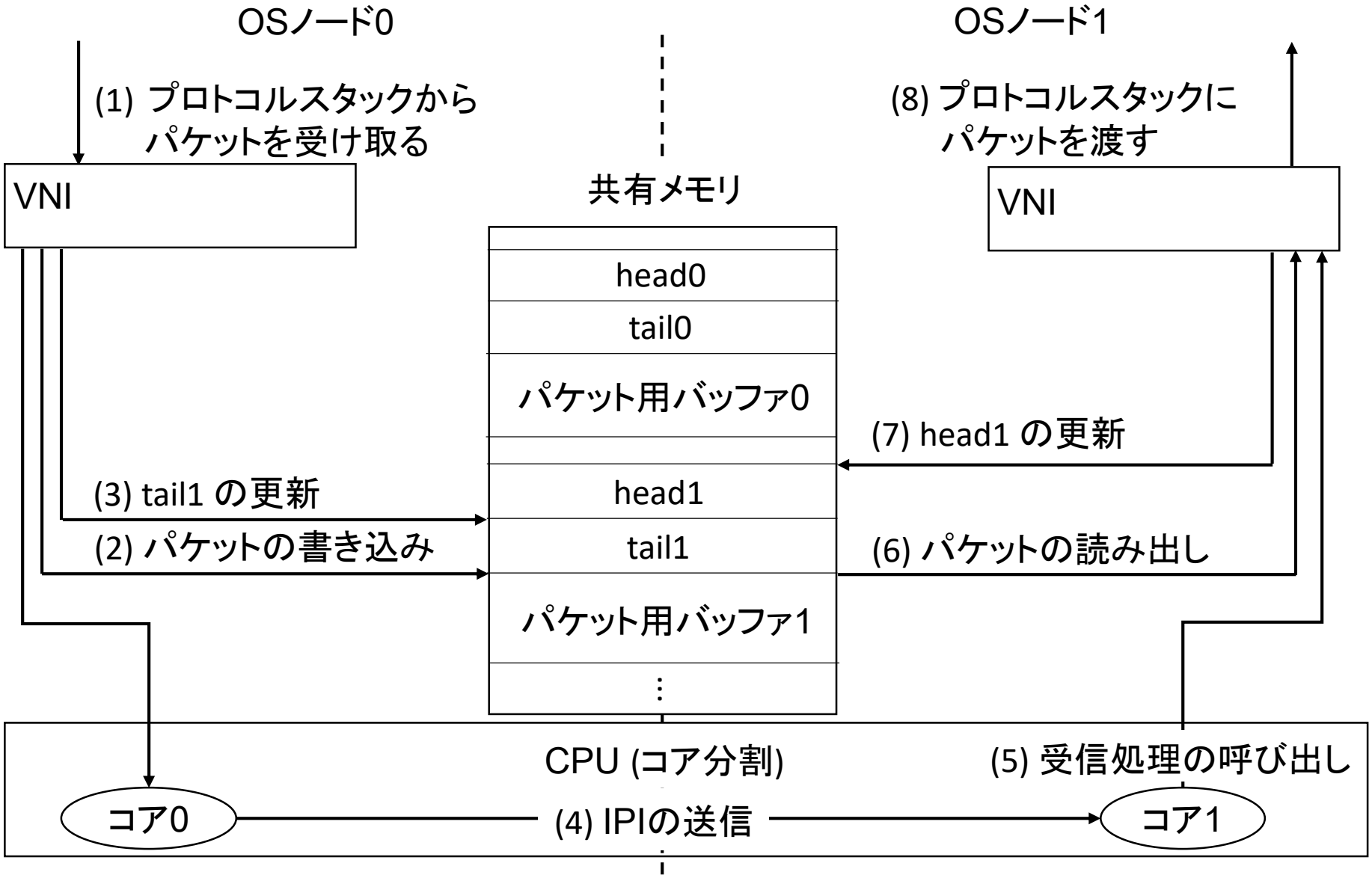
Mintの構成例


AP
AP
AP
OS ノード 0
OS ノード 1
OS ノード 2
CPU (コア分割)
コア 0
コア 1
コア 2
コア 3
メモリ (空間分割)
共有メモリ
OS ノード 0 用
デバイス
OS ノード 1 用
デバイス
OS ノード 2 用
デバイス
H/W

改変前のVNIを用いた通信の処理流れ



改変後のVNIを用いた通信の処理流れ



改変後の送受信バッファにおける問題点

(1) 通信するOSノードの増加に伴い，より多くのパケット用バッファを要する

(2) n個のOSノード間で通信するとき，要するパケット用バッファは ${}_nC_2 \times 2$ 個

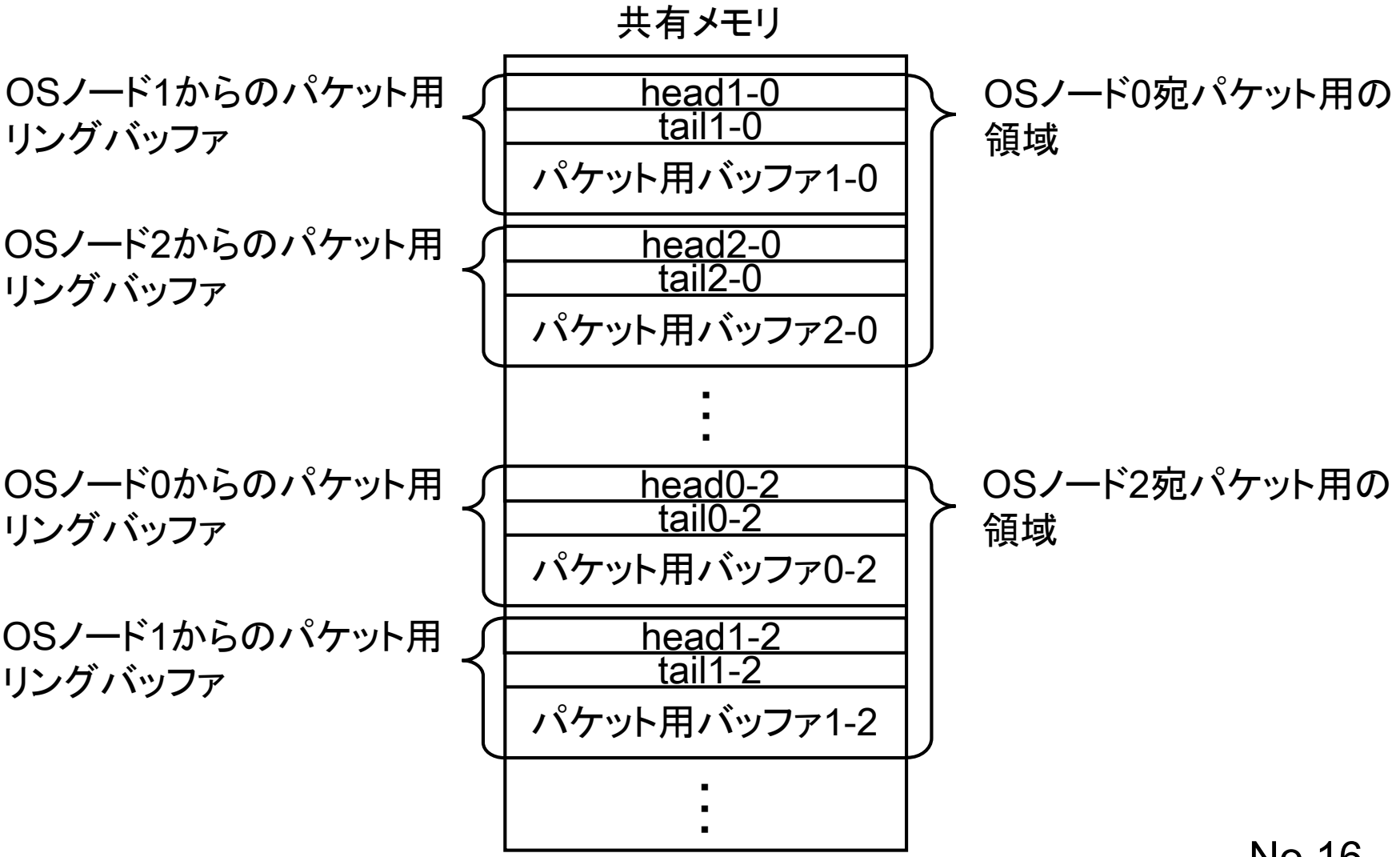
例：通信するOSノード数100のとき

${}_{100}C_2 \times 2 = 4950$

(1) パケット用バッファの大きさの最大値は2つのOSノード間で通信する場合の1/2475

(2) パケット用バッファの大きさを392,200B (計測時と同じ大きさ)とした場合，送受信バッファの大きさは約1.8GB

通信するOSノード数が3のときの送受信バッファ



送受信バッファにおける構成の比較

	利点	欠点
改変前	パケット用バッファの利用効率高	要排他制御 送受信に要する処理多
改変後	排他制御不要 送受信に要する処理少	パケット用バッファの利用効率低

(1) 排他制御の必要性 ・・・改変後が有利

排他制御によるオーバーヘッド，実装の手間に影響

(2) パケット用バッファの利用効率 ・・・改変前が有利

各OSノード間で利用できるパケット用バッファのサイズに影響

(3) 送受信に要する処理の量 ・・・改変後が有利と予想

通信のオーバーヘッドに影響