Technology Consulting Company
Research, Development &
Global Standard

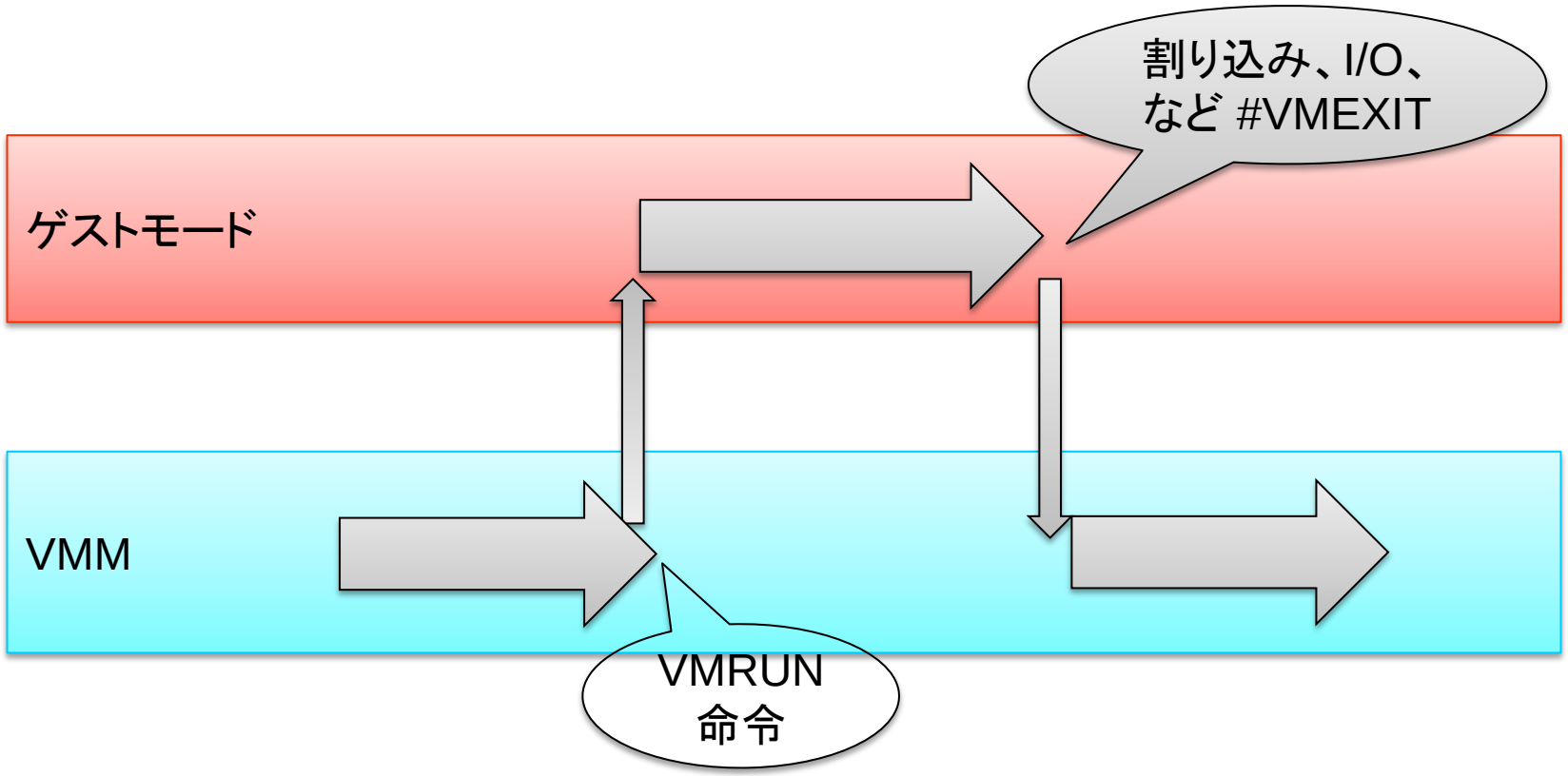
Unsafe Nested Virtualization

榮樂 英樹
株式会社イーゲル

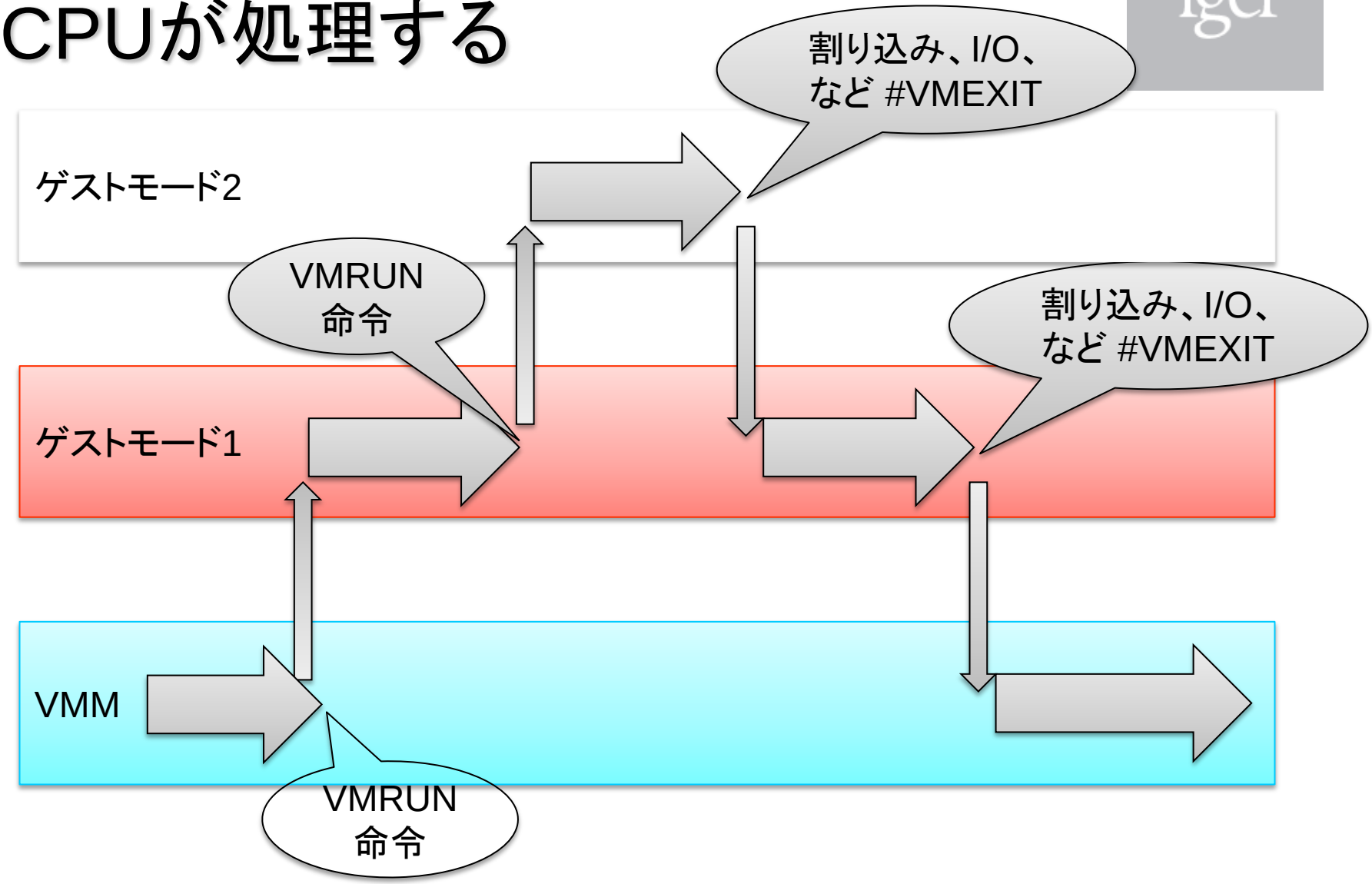
2015-11-26 BitVisor Summit 4

Nested Virtualization

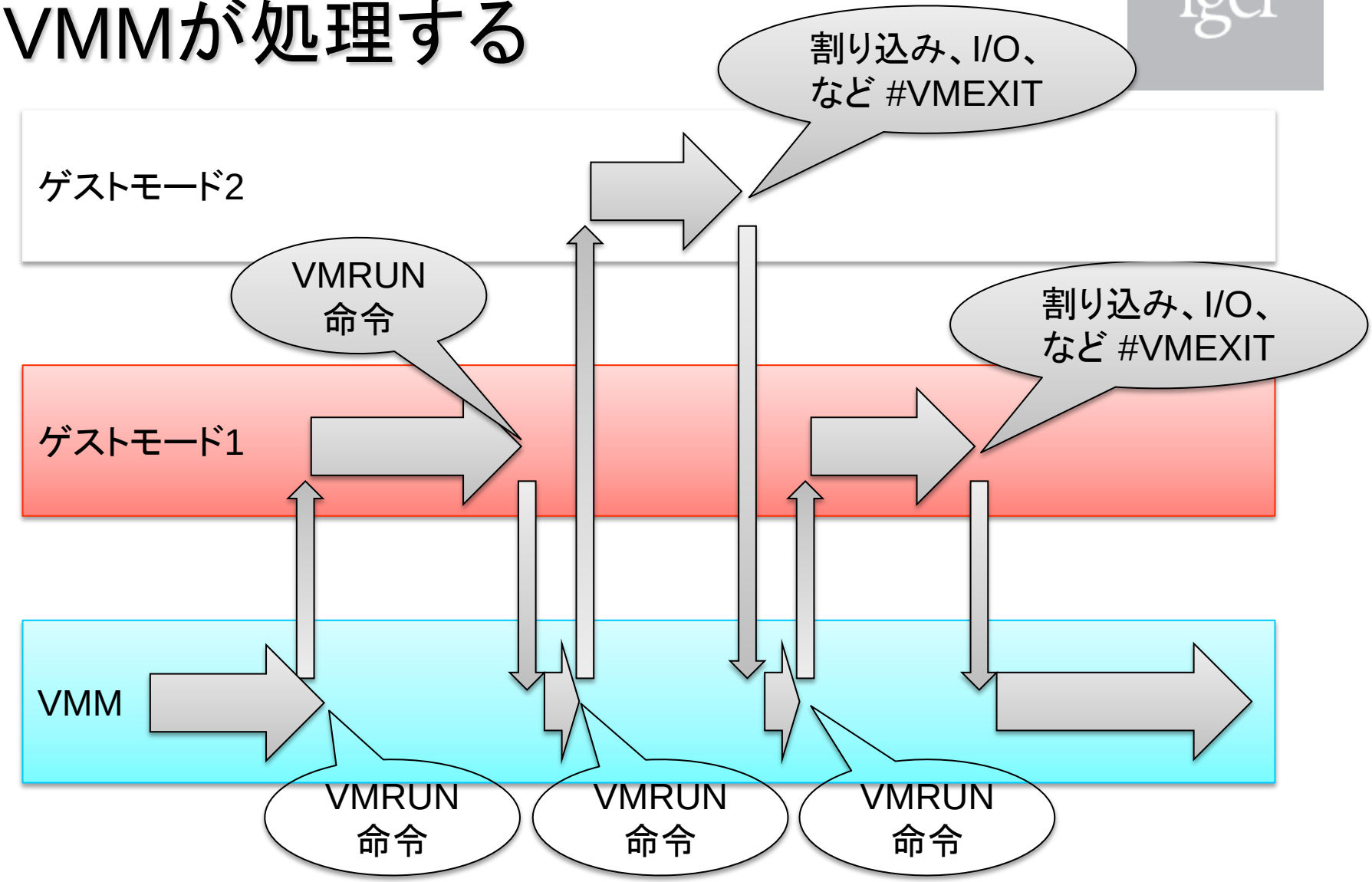
仮想マシン動作時の制御の流れ (AMD SVMの場合)



Nested Virtualizationの理想: CPUが処理する



Nested Virtualizationの現実: VMMが処理する



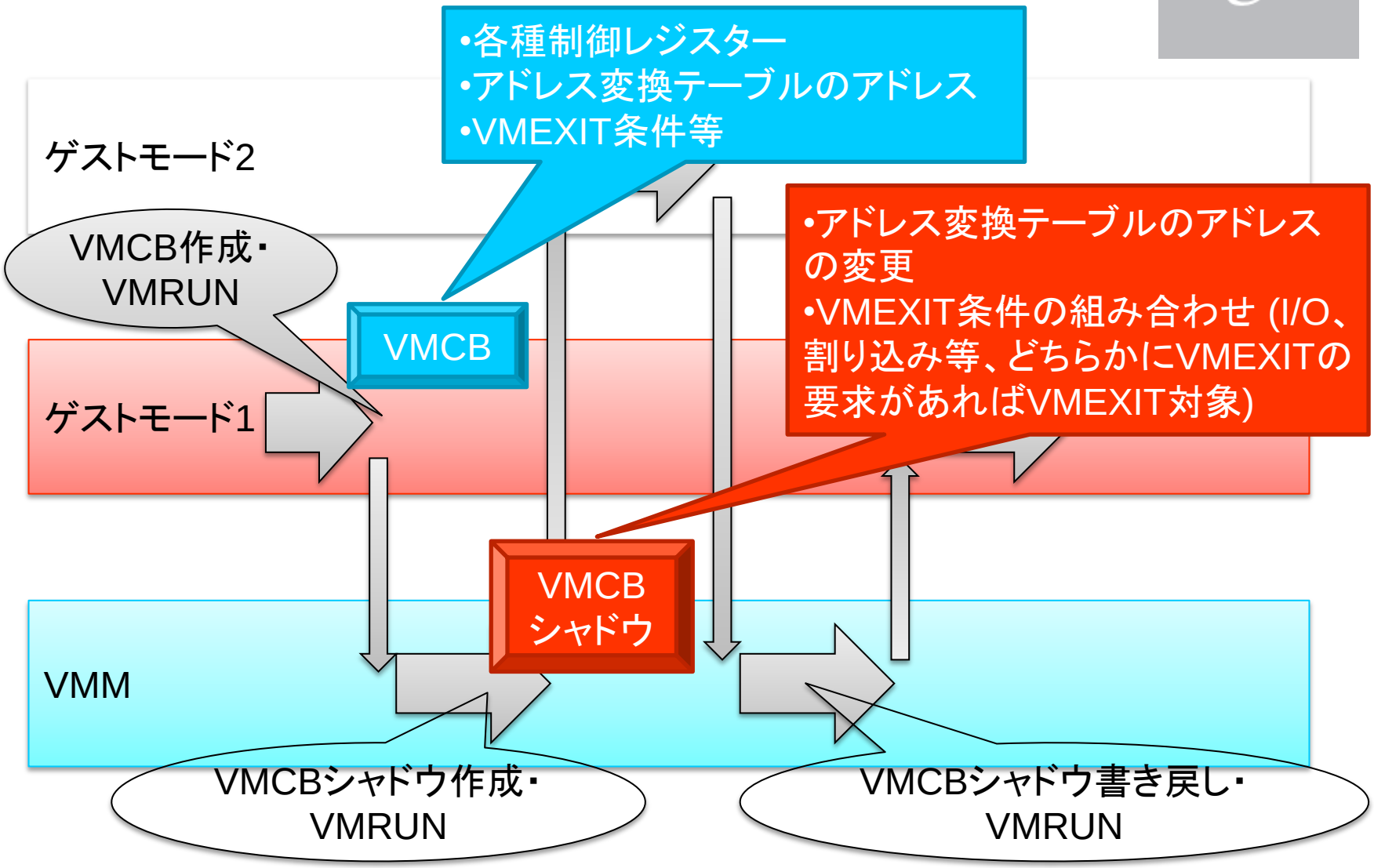
Nested Virtualizationにおける VMMの仕事

igel

- アドレス変換テーブルの管理
- VMCBなどのデータ構造の変換
- VMRUN命令や関連するモデル固有レジスター等の処理

アドレス変換テーブルの管理

VMCBなどのデータ構造の変換



VMCBなどのデータ構造の種類と変換タイミング

igel

データ構造の種類

- VMCS(Intel)/VMCB(AMD)、アドレス変換テーブルのほか、I/Oビットマップ、MSRビットマップなどもある

タイミング

- Intel VT-x: VMCSはVMREAD/VMWRITE命令でアクセスするため、それらのVMEXIT時に変換可能
- AMD SVM: VMCBはRAM上に存在するため、シャドウを作成してVMRUN命令のVMEXIT時にコピー/変換
- その他のデータ構造はシャドウを作成しコピー/変換

VMRUN命令や関連するモデル固有レジスター等の処理

igel

- VMRUN命令 (AMD SVM)
  - VMCBのシャドウの作成・更新
  - シャドウを指定してVMRUN命令実行
  - シャドウから書き戻し

- モデル固有レジスター (以下MSR)
  - 仮想化支援機能の使用許可・禁止、有効・無効の設定
  - 仮想化支援機能で使用するRAMアドレスの設定

Unsafe Nested Virtualization

igel

準パススルー型VMMの特性を活用して大幅に簡略化

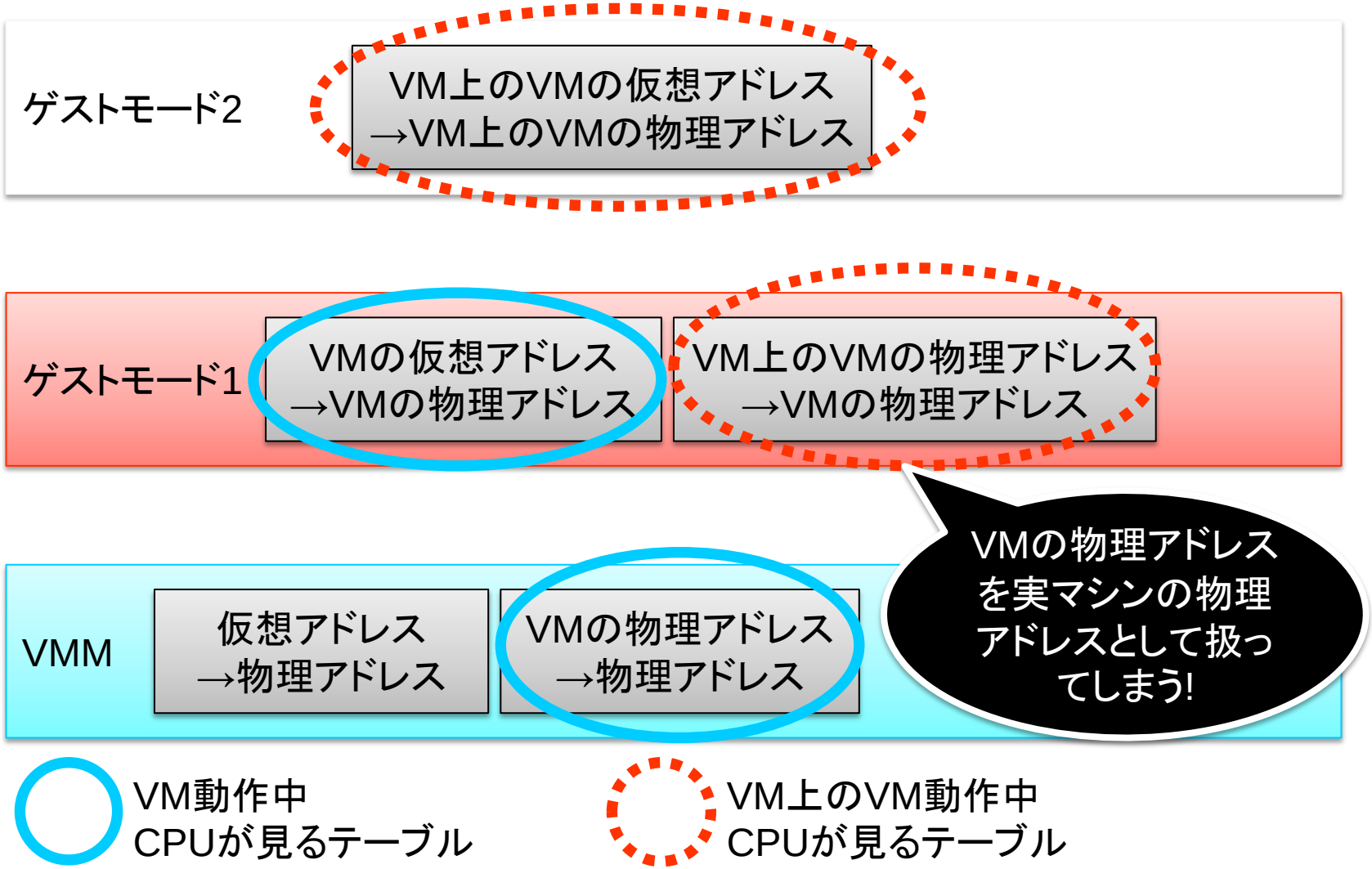
- アドレス変換テーブルの管理の省略
- VMCBなどのデータ構造の変換の省略

Unsafe: VM上のVMが以下のようなことをしないと仮定

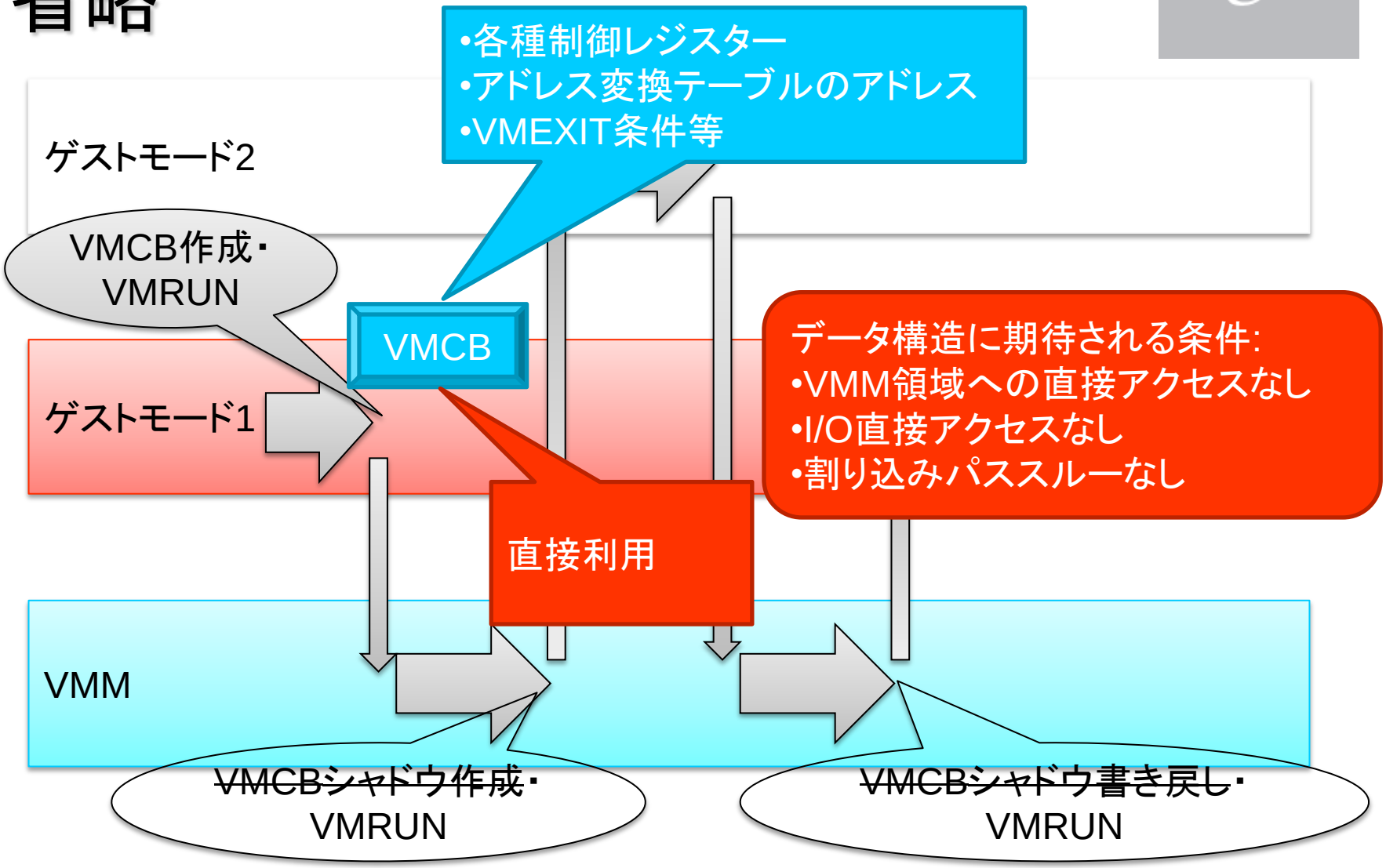
- 暗号化対象デバイスのパススルーアクセス
- VMM領域への直接アクセス
- ホスト側割り込みのパススルー

→VM上のVMMが信頼できれば問題ない

アドレス変換テーブルの管理の省略



VMCBなどのデータ構造の変換の省略



実装

- Intel VT-x - 未着手
- AMD SVM - 実装済み

Intel VT-xでの実装 未着手の理由 (1/2)

igel

- 仮想化支援機能の命令数の違い
  - Intel: VMCALL, VMCLEAR, VMLAUNCH, VMPTRLD, VMPTRST, VMREAD, VMRESUME, VMWRITE, VMXOFF, VMXON
  - AMD: VMRUN, VMMCALL, VMLOAD, VMSAVE, STGI, CLGI

簡単!

- オペランドの違い
  - Intel: 仮想アドレス、複雑なアドレッシング 例: [r8*8+r10+10]
  - AMD: 物理アドレス、レジスター固定

簡単!

Intel VT-xでの実装 未着手の理由 (2/2)

igel

- VMEXIT時の処理の違い
  - Intel: VM entry失敗の場合はVMLAUNCH/VMRESUMEの次の命令から再開、VM exitの場合はVMCSに書かれた命令ポインターから再開
  - AMD: 常にVMRUNの次の命令から再開

簡単!

- アドレス空間IDの違い
  - Intel: VPID - Core i3/i5/i7 世代以降で利用可能、最大65535
  - AMD: ASID - すべてのAMD SVM対応CPUで利用可能、最大NASID-1 (NASIDはCPUIDで取得)

CPU判定不要、ASIDをVMM用に予約可能!

AMD SVMでの実装

igel

- ASID (アドレス空間ID)
- CPUID
- MSR
- #VMEXIT

ASID (アドレス空間ID)

igel

0	1	2	…	NASID-2	NASID-1
VMM	VM上のVM				VM上のVMM

- BitVisorはVM 1つしか使用しないため、ASIDを1つ予約(NASID-1)
- 無駄なTLBフラッシュを回避

- NASID=2の場合はASID=1を利用し、VMRUNのたびにTLBフラッシュを行う
  - 未テスト

CPUID

- AMD SVMを隠ぺいしていたコードを削除
- NASIDをデクリメント
  - NASID=2の場合はデクリメントしない

MSR

igel

- MSR_AMD_VM_CR
  - SVMDIS, LOCK: SVM使用許可・禁止をBIOS・ファームウェアが設定するためのもの、AMD SVM隠ぺいをやめたため、unsafe nested virtualizationが無効の場合は本レジスターでSVM禁止と返す
- MSR_AMD_VM_HSAVE_PA
  - 仮想化支援機能用のデータ領域のアドレス、VMMですでに確保済みのため、設定されたアドレスは実際には使用しない
- MSR_IA32_EFER
  - SVME: SVM有効・無効、VMから設定可能なように見せるが実際には常にセットしておく (最初からセットしておくとLinux KVM等は他のVMMと競合していると判断して動作しない)

#VMEXIT

igel

- VMEXIT_STGI, VMEXIT_CLGI
  - NMIを含めた割り込み許可禁止でVMRUNの前後で使用される
  - V_INTR_MASKINGを使用してNMI割り込み禁止を実現
- VMEXIT_INVLPGA
  - ASIDを指定したTLBフラッシュ
  - ASID=0の場合INVLPG命令と同じ処理を実行
  - それ以外の場合VMMが代理でINVLPGA命令を実行
- VMEXIT_VMRUN
  - NASID=2の場合はTLBフラッシュの設定の書き換えと復元
  - VMRUN命令を代理で実行
  - 割り込み許可フラグはVMの状態を反映

その他の命令

igel

- VMMCALL
  VM上からVMMを呼び出すために使用する命令で、実マシン上のVMMでは使われない (VMMで実行されれば実マシン上では例外が生成されるが、VM上のVMMは実行しないものと仮定)

- VMLOAD, VMSAVE
  VMCBから一部の制御レジスター等をロード・セーブする命令で、物理アドレス指定のため通常はVMMが処理するべきだが、Unsafe Nested Virtualizationなので省略 (命令のinterceptをオフに)

動作確認

- AMD A10-7870K / Ubuntu 64bit
  - VirtualBoxとLinux KVMが軽快に動作
- AMD Z-01 / Windows 7 32bit
  - VirtualBoxが正常に動作せず
  - 原因調査中

まとめ

igel

- Nested Virtualizationの概要
  - アドレス変換テーブルの管理、VMCBなどのデータ構造の変換、VMRUN命令や関連するモデル固有レジスター等の処理
- Unsafe Nested Virtualization
  - 準パススルー型VMMの特性を活用して大幅に簡略化
  - VM上のVMMが信頼できれば問題ない
  - Intel VT-xでの実装未着手の理由
  - AMD SVMでの実装
  - 動作確認