igel

Technology Consulting Company
Research, Development &
Global Standard

BitVisor 2018年の主な変更点

榮樂 英樹
株式会社イーゲル

2018-11-28 BitVisor Summit 7

BitVisor 2018年の主な変更点

igel

- バグ修正
  - NMI関連の問題
  - IA32_TSC_ADJUST MSRの変更未対応
  - シャドウページテーブルの正しくないfast path
  - タスクスイッチ時の命令ポインター更新忘れ
  - XSS-bitmap exiting未初期化
  - 仮想マシン上でBitVisorを動作させる場合の問題
- 機能追加 (*)昨年のBitVisor Summit 6で発表された機能の取り込み
  - PCID対応
  - dbgsh用telnetサーバー追加
  - Realtek NIC＋virtio-net対応
  - NVM Express対応 (*)
  - Intel VT-x用unsafe nested virtualization対応 (*)
- その他細かな修正

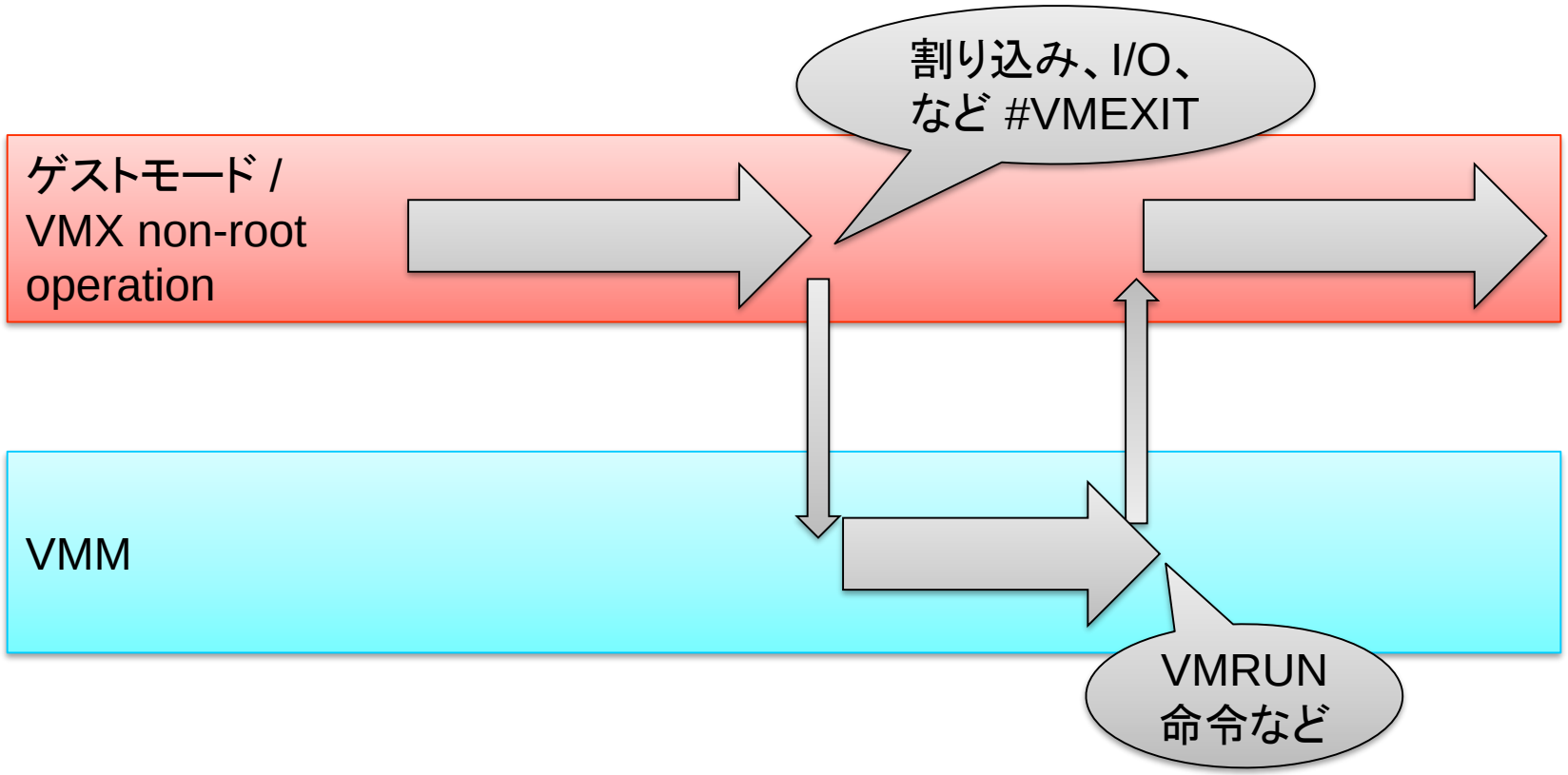
バグ修正: NMI関連の問題

igel

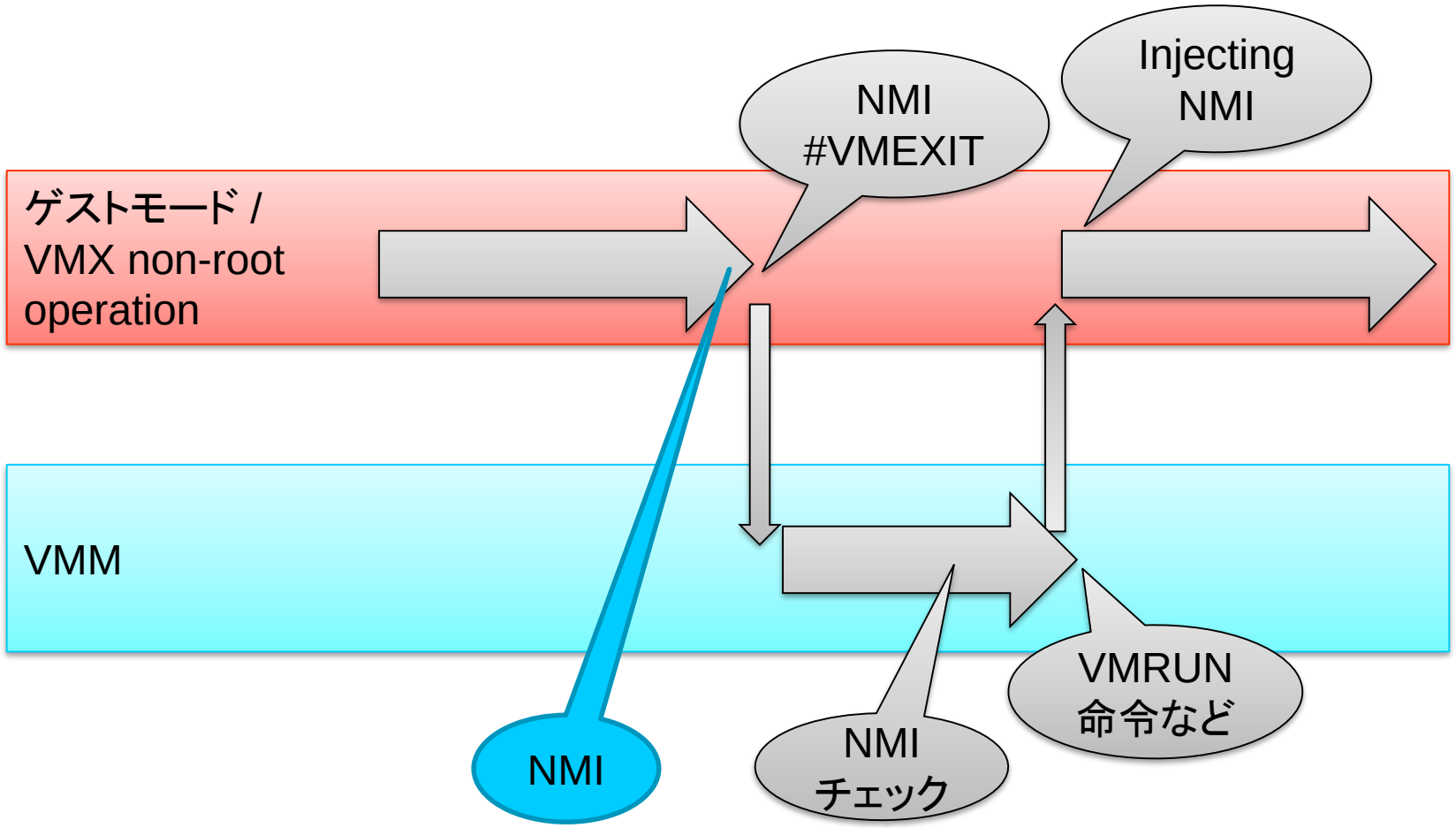
(NMI: Non-Maskable Interrupt)

- NMI生成が遅れる
- NMIブロック中のNMIを正しく処理しない
- NMIハンドラーがプロセス（保護ドメイン）の%gs変更を想定していない
- NMIがプロセスのシステムコールの直後に発生することを想定していない

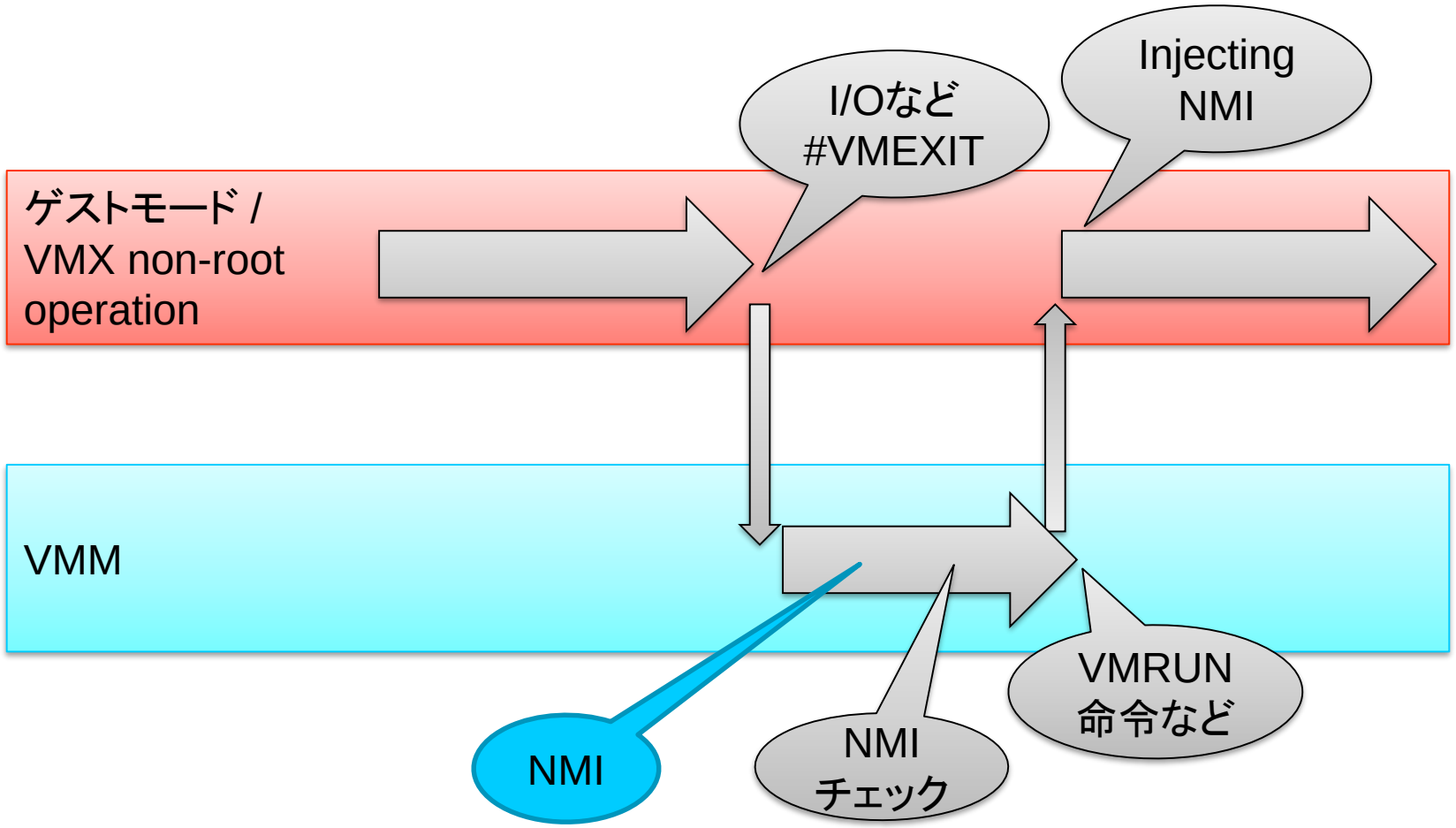
仮想マシン動作時の制御の流れ



今までもOKだったNMI
(1) NMIによる#VMEXIT

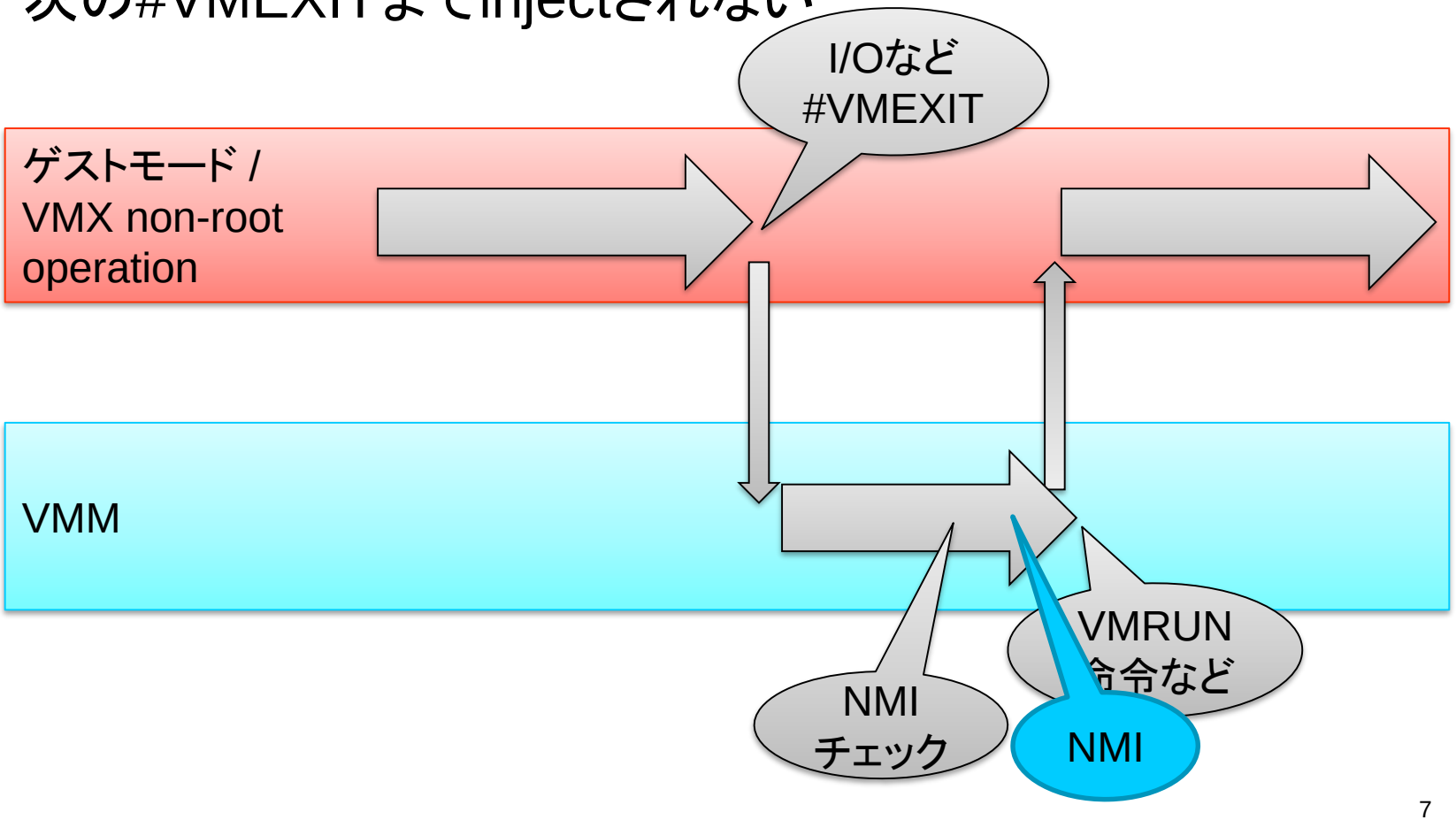


今までもOKだったNMI
(2) NMIチェック前のNMI

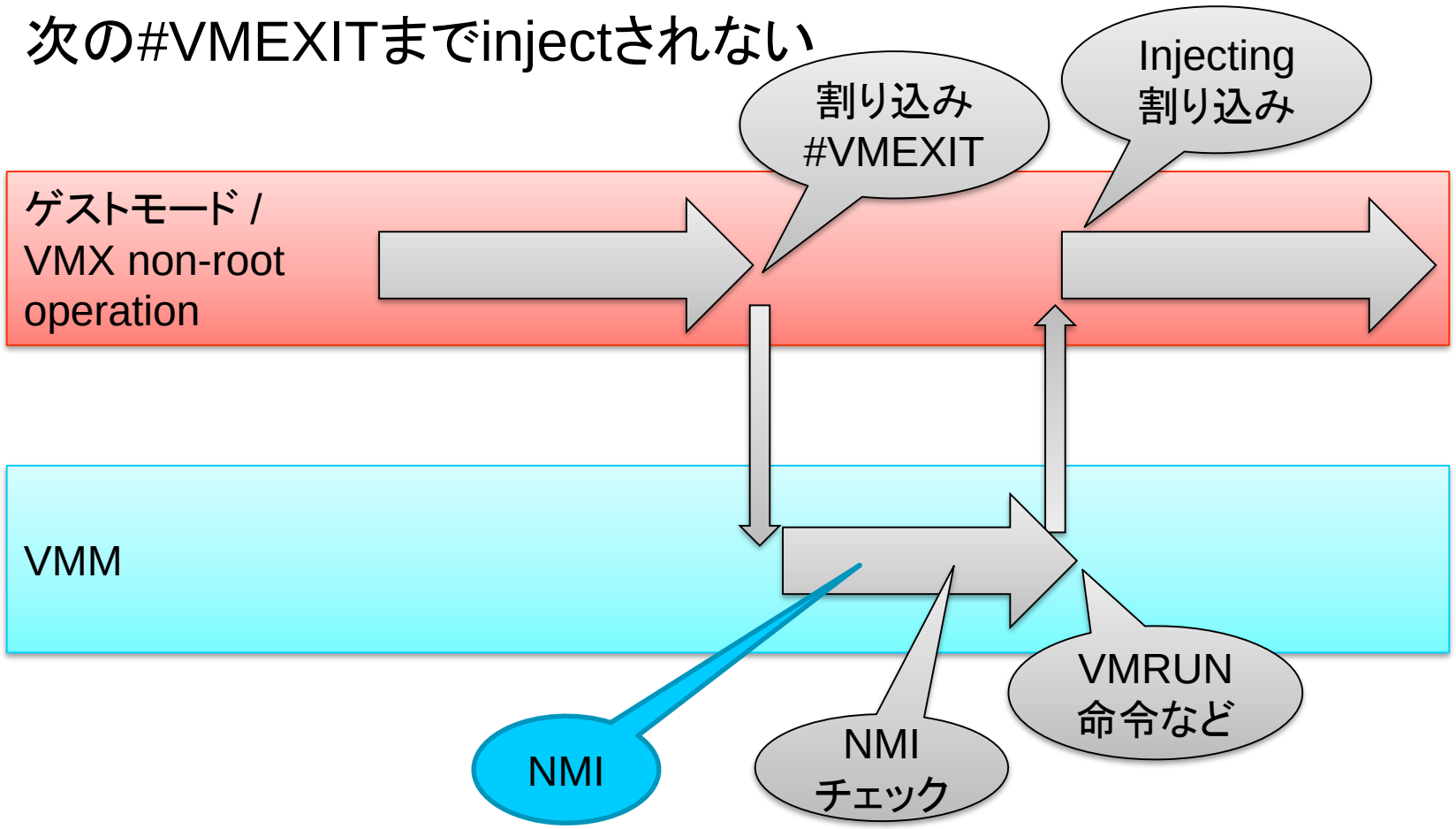


NMI生成が遅れる問題
(1) NMIチェック後のNMI

次の#VMEXITまでinjectされない



NMI生成が遅れる問題
(2) Event Injectionの競合



NMI生成が遅れる問題の具体例: Windows再起動でのハングアップ

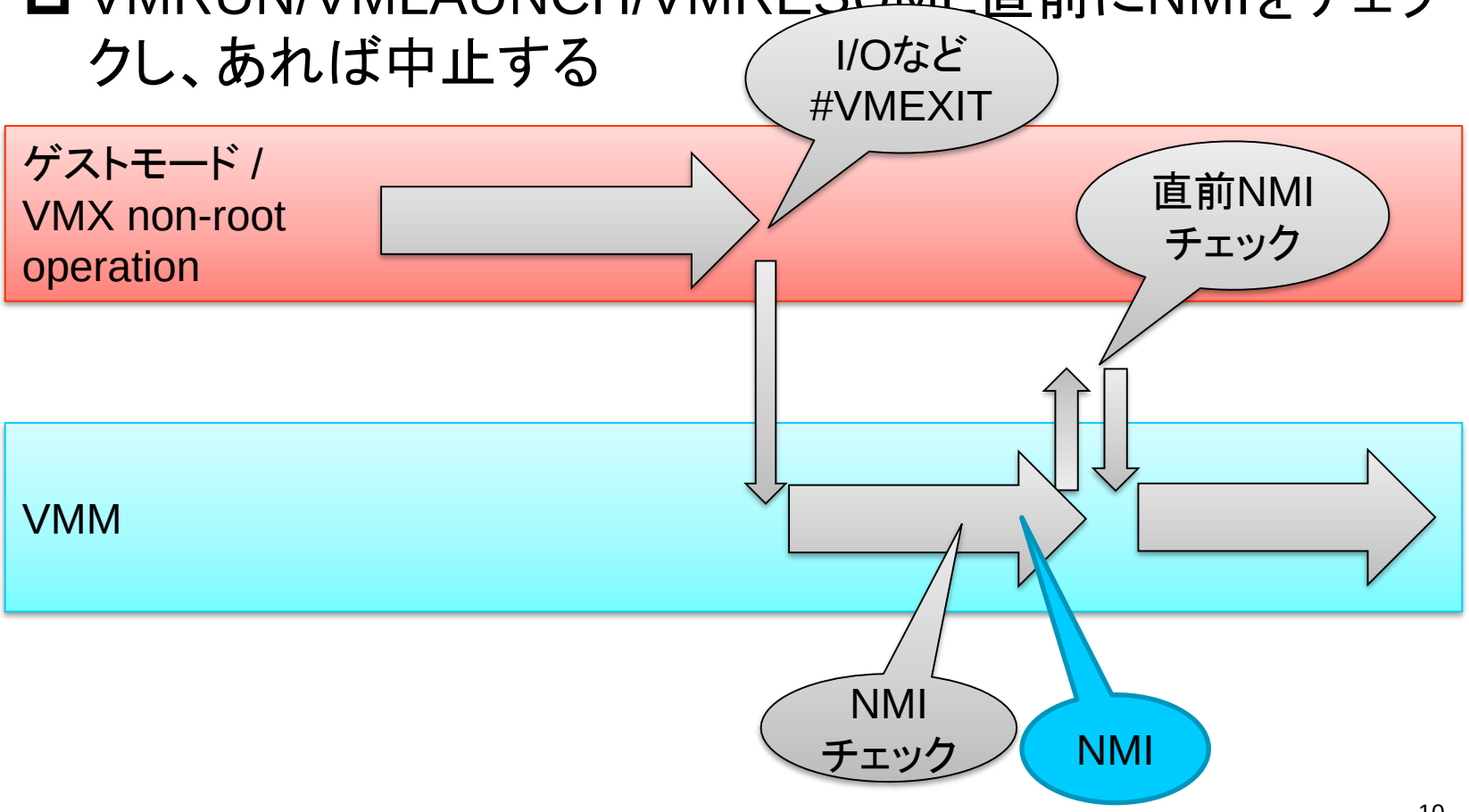
igel

1. Windowsが再起動の際にCPU0からNMIを送信し他のプロセッサの応答を待つ
2. CPU1がたまたま割り込み#VMEXITの処理中でNMI生成が遅れる
3. CPU1のWindowsが、割り込み処理後（おそらく）NMIを待つため割り込み禁止にしてMWAIT命令を実行する
4. CPU1は二度と#VMEXITすることはなく動作を停止する
5. CPU0はCPU1の応答を待ち続け無限ループに陥る

NMI生成が遅れる問題の対策
(1) NMIチェック後のNMI

igel

- VMRUN/VMLAUNCH/VMRESUME直前にNMIをチェックし、あれば中止する



VMRUN/VMLAUNCH/VMRESUME直前にNMIをチェック

igel

競合を避ける必要がある。

AMD SVMの場合:

- CLGI命令でNMIをマスク(?)してチェックする
- マスク中にNMIが発生した場合は、VMRUNでマスク解除されるためその直後に#VMEXITとなる

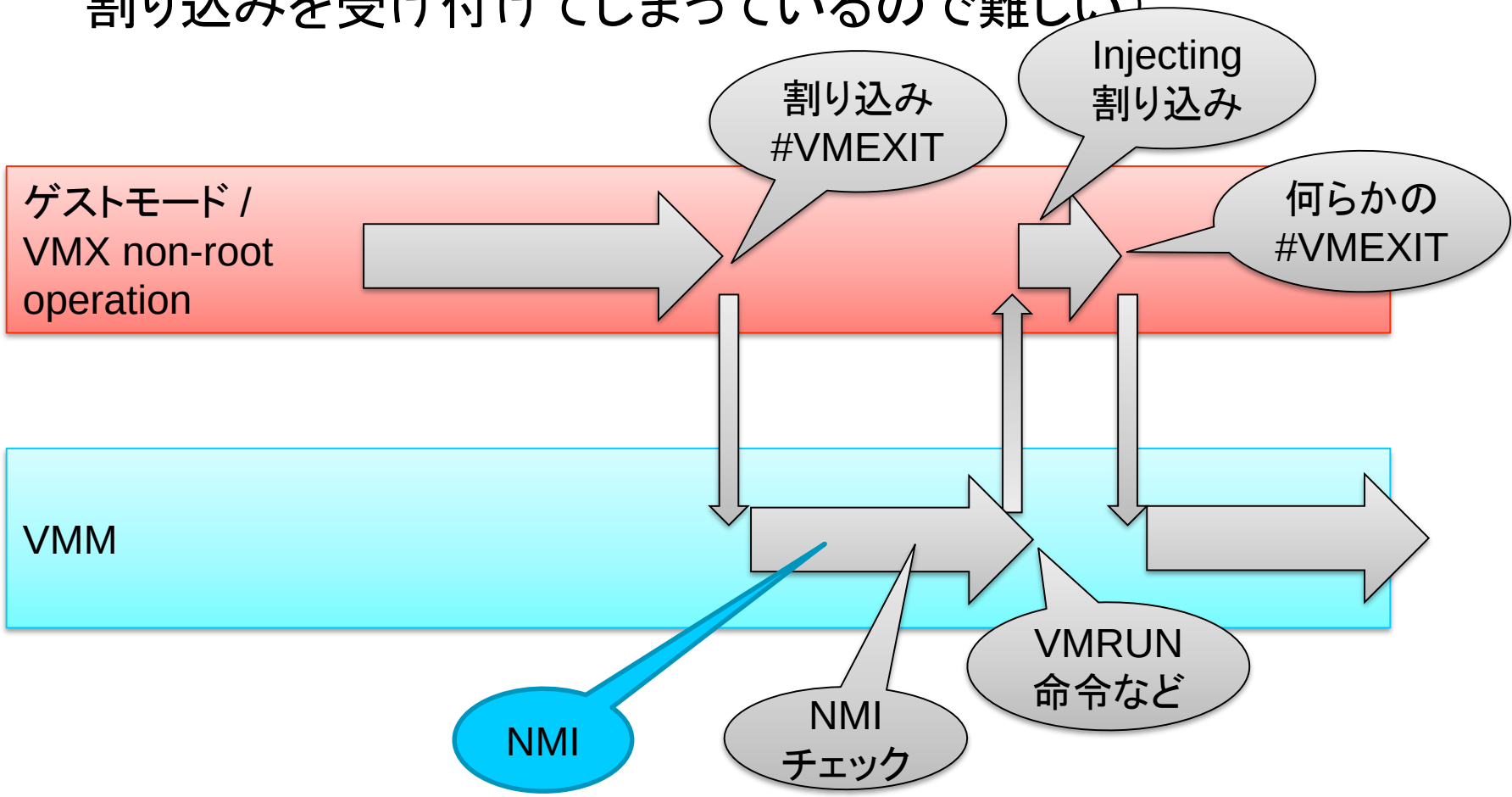
Intel VT-xの場合:

- NMIハンドラー内で、スタック上のリターンアドレスを見て、VMLAUNCH/VMRESUME直前のアドレス範囲の場合はNMI用のアドレスに書き換えるようにする
- そのアドレス範囲内でNMIをチェックする

NMI生成が遅れる問題の対策
(2) Event Injectionの競合

igel

- 直後に#VMEXITさせる（本来NMIの優先度が高いが、割り込みを受け付けてしまっているので難しい）



直後に#VMEXITさせる方法

igel

AMD SVMの場合

- 良い方法がなかったので、HLT/MWAIT/PAUSEの各命令のinterceptを有効にして、Windowsの具体例については対策する

Intel VT-xの場合

- NMI-window exitingを使用する
  - NMI生成可能な状況になればすぐにVM exitされる

NMIブロック中のNMIを正しく処理しない問題

igel

NMIハンドラーが呼び出されてからIRET命令が実行されるまでに発生したNMIはブロックされる（80286以降）。ここではNMIハンドラーが呼び出されてからIRET命令が実行されるまでを**NMIブロック中**と表す。

これまでのBitVisorの動作:

- Intel VT-x: NMIブロック中のNMIは失われる（IRET後も生成されない）
- AMD SVM: NMIブロック中のNMIは生成される（もしNMIハンドラーとしてタスクスイッチが設定されていれば不正なタスクスイッチで例外発生の原因に）

NMIブロック中のNMIを正しく処理しない問題の対策（Intel VT-x）

igel

- VMCS内にBlocking by NMIというビットがあり、NMIブロック中はこれがセットされている
- これまでBlocking by NMIがセットされているとNMIを破棄していた
- Blocking by NMIがセットされているときにNMIを破棄するのではなく、NMI-window exitingをセットすることで、IRET命令の直後にVM exitさせて、NMIを生成することができる

NMIブロック中のNMIを正しく処理しない問題の対策（AMD SVM）

igel

- NMIブロック中はIRET命令のinterceptを有効にし、それが有効の間はNMI生成をブロックする
  - IRET命令の#VMEXITでブロックを解除したいが、実際にはIRETの次の命令までは次のNMIは生成できない
  - しかしIRET命令のエミュレーションは面倒（特にタスクスイッチ）
- そのためIRET命令時点でinterceptを無効にしたうえで、IRET命令で外部割り込みが発生するのを避けるためinterrupt shadowをセットし、NMI生成が必要な場合はHLT/MWAIT/PAUSEの各命令のinterceptを有効にしてお茶を濁す

NMIハンドラーがプロセスの%gs変更を想定していない

igel

- %gsをいきなり使っていたため、プロセス（保護ドメイン）が別のセグメントをロードすれば不正なアドレスへのアクセスをすることになっていた
- %gsをセットしてから使うようにして解決

NMIがプロセスのシステムコールの直後に発生することを想定していない

igel

- 64bitではSYSCALL命令によるシステムコールを使用していた
- SYSCALL命令はスタックポインターを切り替えない
- BitVisorはNMIハンドラーがスタックを切り替えるよう設定していない
- SYSCALL命令の直後にNMIが発生するとプロセスのスタックのままNMIハンドラーが実行されるおそれがあった
- 64bitコールゲート対応を実装し、SYSCALLはデフォルトでオフになるようにした（CONFIG_USE_SYSCALL64）
  - QEMUでは64bitコールゲートが実装されていないらしく、エラーが出るようになった...

バグ修正: IA32_TSC_ADJUST MSRの変更未対応

igel

- IA32_TIME_STAMP_COUNTER MSR: TSCそのもの
- **IA32_TSC_ADJUST MSR**: 基準となるTSCとの差

BitVisorの処理:

- IA32_TIME_STAMP_COUNTER MSRの変更: 実際のTSCは変更せず、その時点のTSCとの差をTSC offsetにセットする
- **IA32_TSC_ADJUST MSR**の変更: 実際のTSCは変更せず、与えられた値をTSC offsetにセットする
  - この実装がなくパススルーされてしまったために実際のTSCが変更されてしまい、BitVisor内の時刻情報がおかしくなっていた

バグ修正: シャドウページテーブルの正しくないfast path

igel

- CONFIG_CPU_MMU_SPT_3（デフォルト）使用時、ゲストOSがCR0・CR3・CR4・EFERの各レジスターを変更せずにTLBフラッシュした場合、PTE・PDEの変更のあった部分だけを反映するfast pathが実装されていた
- このfast pathではPDPTE・PML4Eの変更が反映されず、再利用されたため正常にTLBフラッシュできていなかった
- 正しくないfast pathだったので削除した

見つかった経緯: BitVisor上でBitVisorを起動してNMI問題の実験中、たまたまCore 2 DuoのPCを使用したところプロセスが異常終了したことで発覚。BitVisorのプロセスの空間はCR3ではなくPDPTEの変更により切り替えられていたため、完全に上の条件を満たしていた。

バグ修正: タスクスイッチ時の命令ポインター更新忘れ

igel

- タスクスイッチ時に命令ポインターを進めていないバグがNMI問題の調査中に発見された
  - タスクスイッチの種類: JMP, CALL, 割り込み、IRET
  - 命令ポインターを進めてはいけないもの: 外部割り込み、NMI, INT3・INTO以外の例外
  - 命令ポインターを進めなければならないもの: JMP, CALL, ソフトウェア割り込み（INT）、INT3・INTOによる例外, IRET
- 修正した
  - Intel VT-x: 命令長はVMCSから取得。INT3・INTOは「ソフトウェア例外」として区別。
  - AMD SVM: 命令長は機械語を読み取って解釈。INT3・INTOは例外の割り込み番号により区別。

バグ修正: XSS-bitmap exiting未初期化

igel

- XSAVES/XRSTORS命令対応の際に、XSS-bitmap exitingを初期化していなかったことが分かった
  - Windowsのシャットダウン時にXSAVES命令によるVM exitでpanicが発生したことで発覚

- XSS-bitmap exitingを0クリアするようにした

バグ修正: 仮想マシン上でBitVisorを動作させる場合の問題

igel

- On Linux KVM
  - KVMの拡張機能による予期しないページフォールト
  - 外部割り込み生成時のVM entry failed panic

KVMの拡張機能による予期しないページフォールト

igel

- Linux KVMの拡張機能でasynchronous page faultというのがあって、その機能によりページフォールトがBitVisor内の予期しないタイミングで発生してしまうことが判明した
  - BitVisor内の外部割り込み受付処理

    ```
    sti
    nop
    cli    ; ここでページフォールトが発生
    ```

  - BitVisorはこれを**外部割り込み**の14が発生したものとして処理してしまい、オペレーティングシステムが正常に動作できなくなる
- CPUIDでハイパーバイザーの拡張用に使われる0x40000000-0x4FFFFFFFを隠ぺいすることにより解決

外部割り込み生成時のVM entry failed panic (1/2)

igel

- BitVisorが外部割り込みをinjectしようとしたときに、Intel VT-xのblocking by STIがセットされているとVM entry failed panicが発生する (on Linux KVM)

  Blocking by STIについて:

  ```
  sti         ; ここで割り込みを許可
  nop         ; ここでは割り込みは発生しない (Blocking by STI)
  cli
  ```

- 今まで発覚しなかったのはなぜか
  - BitVisorは割り込みかinterrupt windowのVM exitで割り込みをinjectしていた
  - Interrupt windowはblocking by STIでは発生しない仕様である
  - 実際のCPUは、おそらくblocking by STIの状態では割り込みによるVM exitをしないのではないか

外部割り込み生成時のVM entry failed panic (2/2)

igel

- Blocking by STIチェックのために、割り込みVM exitに関する処理を書き直した
  - 仮想マシンがパススルーかどうかを示すフラグを追加し、VT-x/SVM実装側にパススルー処理を実装した

Intel VT-x

- 割り込みVM exit: 仮想マシンの割り込み許可状態にかかわらず発生
  - Interrupt-window exitingをセットする
  - Blocking等の条件が合えばすぐに割り込みをinjectする
- Interrupt window VM exit
  - 割り込みをinjectする

AMD SVM

- 割り込み#VMEXIT: 仮想マシンが割り込み許可状態で割り込みを受け付ける直前に発生
  - すぐに割り込みをinjectする

機能追加: PCID対応

Intel CPUにおけるMeltdown脆弱性（CVE-2017-5754）の対策において、性能低下を抑える目的でPCIDを活用するオペレーティングシステムが2018年に急増したため、BitVisorのPCID対応を行った

- EPTとUnrestricted guestが有効、かつ、PCIDが使用可能な時に限り、PCID使用可能とし、INVPCIDも対応していれば使用可能とした
- 今のところ、シャドウページテーブルを使う可能性がある環境と、AMD SVMでは使用可能にしていない

機能追加: dbgsh用telnet サーバー追加

igel

- lwIPを用いた簡単なtelnetサーバーを実装し、それを用いて遠隔からdbgshにアクセスできる機能を追加した
- 有効化方法:
  - vmm.telnet_dbgsh=1
- 有効化されていれば、panicの際にもシェル起動の代わりにtelnet modeと出力してネットワーク通信ができる状態で停止する
  - ネットワークデバイスとlwIPが使用可能でなければ何もできなくなる
- 複数セッションには対応せず、二つ目の接続があれば最初の接続は切断される

機能追加: Realtek NIC＋ virtio-net対応

igel

- FreeBSD用Realtek NICドライバーをBitVisorに移植し、virtio-net化して使用できるようになった
- 有効化方法:
  - CONFIG_NET_RE=1
  - vmm.driver.pci=driver=re, virtio=1
    - tty=1もつけるとネットワークログ出力機能が使用可能
    - net=ippass等の指定によりlwIP等が使用可能
- OpenSSLなどと同様に、移植したモジュールはライセンスが異なるため、make configで切り離してビルドできるようにして、デフォルトではオフにしてある

機能追加: NVM Express対応

igel

- 昨年のBitVisor Summit 6で発表された機能の取り込み
- 有効化方法:
  - 通常のNVMeデバイス用:
    vmm.driver.pci=driver=nvme
  - 一部コンピューター向けNVMeデバイス用:
    vmm.driver.pci=driver=nvme_apple

機能追加: Intel VT-x用Unsafe Nested Virtualization対応

igel

- 昨年のBitVisor Summit 6で発表された機能の取り込み
- 一部改良や問題対策等の修正を行った

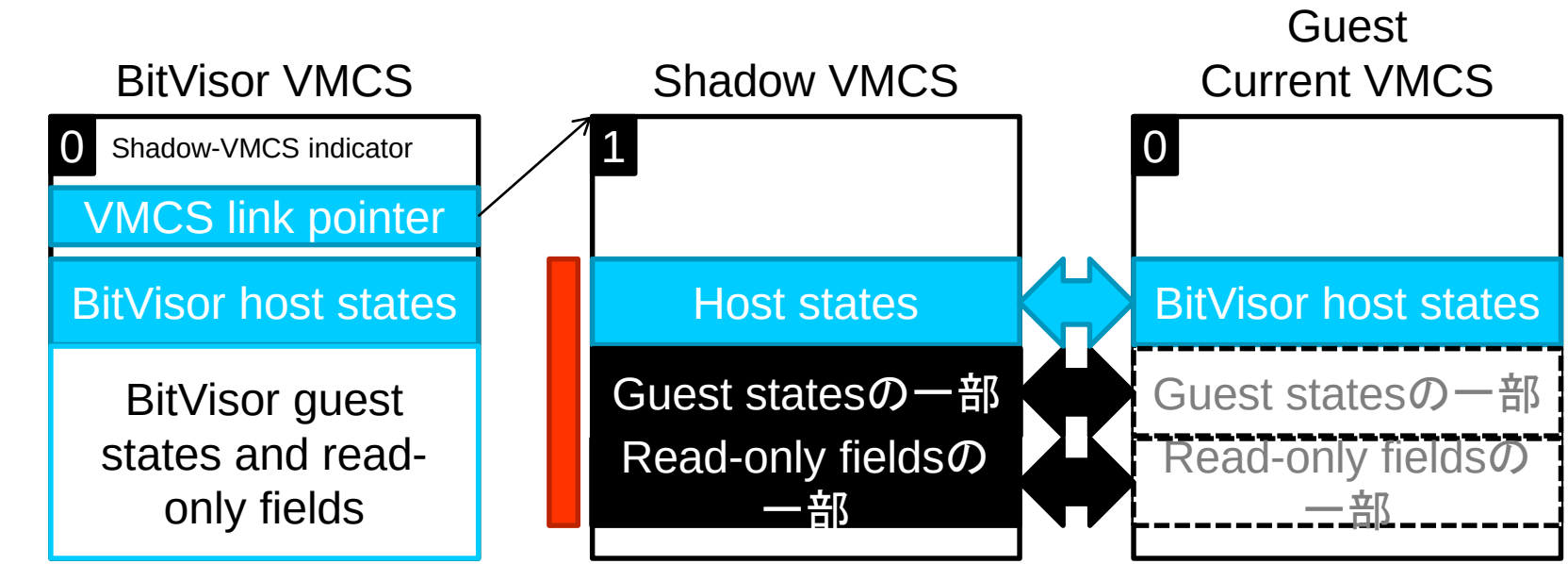
Intel VT-x用Unsafe Nested Virtualization対応 修正内容概要

igel

- **VMCS shadowing対応**
- **Intel NIC＋virtio-netの割り込み番号変換対応**
- VPIDはBitVisorのVMだけが使い、他はVPID=0を使う
- VMFUNC機能隠ぺい
- バグ修正
  - VMCLEAR, VMPTRLD, INVEPT, INVVPID, VMWRITE, VMREADおよびVMXOFF命令のエラーを示すフラグの設定
  - Host statesのセグメント周りの反映処理
  - EFERをVMCSに設定できないCore 2 Duoでのpanic
  - 抜けていた処理: VMPTRST命令、CPLチェック、リセット処理、VMLAUNCHまたはVMRESUME命令でのMOV SSエラーとVM entry failureエラーを返す処理、NMI対応

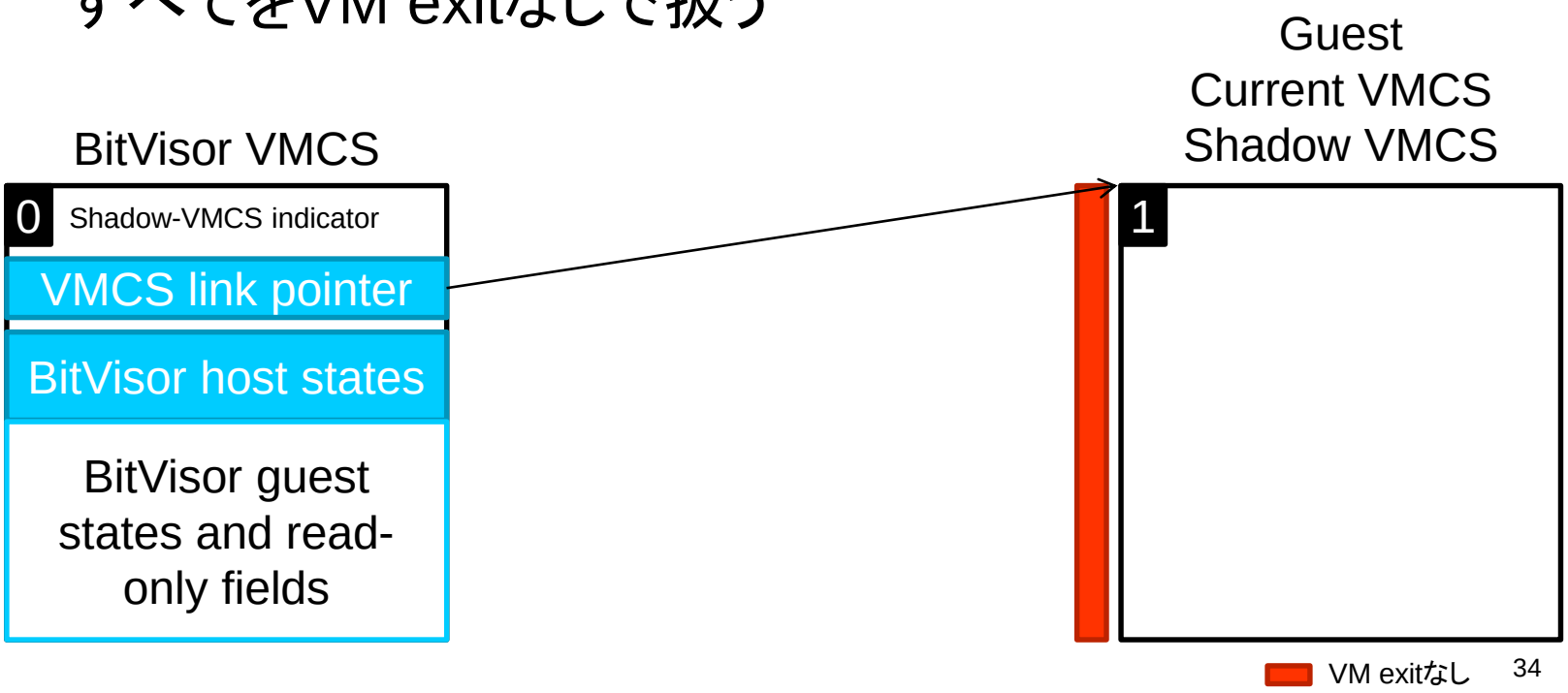
VMCS Shadowing対応

- VMREAD/VMWRITEによるVM exitを減らし高速化
- Host states、よく使われるguest statesと、可能であればread-only fieldsをshadowに保持し、VM entry/exitの際に必要なものをコピー



VMCS Shadowing対応（仮想マシンに対して）

- 仮想マシン上でもVMCS shadowingが使われる場合、shadow-VMCS indicatorが1になっているVMCSは、VMLAUNCH/VMRESUMEに使われることはないため、すべてをVM exitなしで扱う



Intel NIC＋virtio-netの割り込み番号変換対応 (1/2)

igel

- Intel VT-xの "Acknowledge interrupt on exit" 機能
  - 無効の時は、割り込みでVM exitした後、割り込み許可にした時に、割り込み番号に対応する割り込みハンドラーが呼び出される
  - 有効の時は、割り込みでVM exitした時に、VM-exit interruption informationに割り込み番号が入る
  - Linux KVMはこれを有効にしている
- Intel NIC＋virtio-net使用時に影響
  - Intel NIC＋virtio-net機能はMSIで割り込み番号16～31を使用し、仮想マシンに対して割り込み番号を変換している
  - KVM上のVM動作中に割り込み16が発生すると、BitVisorに戻ってきた時点で割り込み16は受付済みであり、放っておくとKVMが割り込み16を生成してfpu exceptionになってしまった

Intel NIC＋virtio-netの割り込み番号変換対応 (2/2)

igel

- VMCSのVM-exit interruption informationに含まれる割り込み番号を書き換える必要がある
  - Read-only fieldsのひとつで、古いCPUでは書き換えができない
  - 書き換えができる場合は素直に書き換える
  - 書き換えができない場合は値をとっておき次のVMREADの時に正しい値を返す
    - VMPTRLDで別のVMCSに切り替えられた場合や、VMCLEARが使用された場合には対応できない
    - 割り込みVM exitだからすぐに読み取って処理してくれるだろうという淡い期待に基づいている
- 割り込みをなかったことにすることは難しい
  - とりあえずVM-exit interruption informationのvalidビットをクリアするようにしたが、矛盾が生じている

その他細かな修正

igel

- INIT signal処理のリファクタリング
- リセット処理で設定し忘れていたステートやレジスターの設定
- strcmp/memcmpでsetnc/setc/rorの代わりにsbb/xorを使用
- mapmem失敗時のアンロック
- 32bitビルド版でmapmem領域を増やした（PCI MMCONFIGのマップ失敗対策）
- 32bitビルドエラー修正
- gs_*定義削除
- cpu_seg_*()関数をlong modeでも使用可能に
- VT-x: DEBUGCTL MSRのアクセスをVMCSに反映
- lwIPを2.0.3に更新
- Intel GbE NIC: リンクアップ時のみ送信、デバイス15e3/156f追加
- bnx hotplugパススルー機能追加
- debug dump長さ指定機能追加
- strncmp()追加
- SMX隠ぺい
- 新Mac対応
- ストレージの暗号化とI/Oの分離
- EFERを64bit幅でアクセス
- UEFI: リンク時--subsystem使用
- UEFI: DisconnectController改良
- SVM: SVM-Lock隠ぺい / unsafe nested virtualizationのconfigチェックをMSR処理内で行う

BitVisor 2018年の主な変更点まとめ

igel

- バグ修正
  - NMI関連の問題
  - IA32_TSC_ADJUST MSRの変更未対応
  - シャドウページテーブルの正しくないfast path
  - タスクスイッチ時の命令ポインター更新忘れ
  - XSS-bitmap exiting未初期化
  - 仮想マシン上でBitVisorを動作させる場合の問題
- 機能追加 (*)昨年のBitVisor Summit 6で発表された機能の取り込み
  - PCID対応
  - dbgsh用telnetサーバー追加
  - Realtek NIC＋virtio-net対応
  - NVM Express対応 (*)
  - Intel VT-x用unsafe nested virtualization対応 (*)
- その他細かな修正