Xen – Virtuální hrátky v praxi

Michal Švamberg, Štěpán Kadlec

22. května 2006

Typy virtualizace

Virtualizaci lze rozdělit do tří skupin:

plná virtualizace simuluje kompletní stroj, lze simulovat libovolnou architekturu, je ale vykoupena nejvyšší režijí.
VMware, VirtualPC, QEMU

nativní virtualizace rozděluje stroj na několik menších, minimální režije, ale nízká flexibilita
Virtuozo, Vservers, VirtualPC

paravirtualizace umožňuje běh více hostovaných systémů na speciální architektuře. Architektura Xen/x86 je velmi blízká běžné x86.
Xen, UML

Typy virtualizace

Virtualizaci lze rozdělit do tří skupin:

plná virtualizace simuluje kompletní stroj, lze simulovat libovolnou architekturu, je ale vykoupena nejvyšší režijí.
VMware, VirtualPC, QEMU

nativní virtualizace rozděluje stroj na několik menších, minimální režije, ale nízká flexibilita
Virtuozo, Vservers, VirtualPC

paravirtualizace umožňuje běh více hostovaných systémů na speciální architektuře. Architektura Xen/x86 je velmi blízká běžné x86.
Xen, UML

Typy virtualizace

Virtualizaci lze rozdělit do tří skupin:

plná virtualizace simuluje kompletní stroj, lze simulovat libovolnou architekturu, je ale vykoupena nejvyšší režijí.
VMware, VirtualPC, QEMU

nativní virtualizace rozděluje stroj na několik menších, minimální režije, ale nízká flexibilita
Virtuozo, Vservers, VirtualPC

paravirtualizace umožňuje běh více hostovaných systémů na speciální architektuře. Architektura Xen/x86 je velmi blízká běžné x86.
Xen, UML

Kdo je Xen

Monitor virtuálních strojů:

- velmi dynamicky se rozvíjející projekt
- stojí za ním mnohé významné firmy z IT
- je vyvíjen jako GNU

Kdo je Xen

Monitor virtuálních strojů:

- velmi dynamicky se rozvíjející projekt
- stojí za ním mnohé významné firmy z IT
- je vyvíjen jako GNU

Kdo je Xen

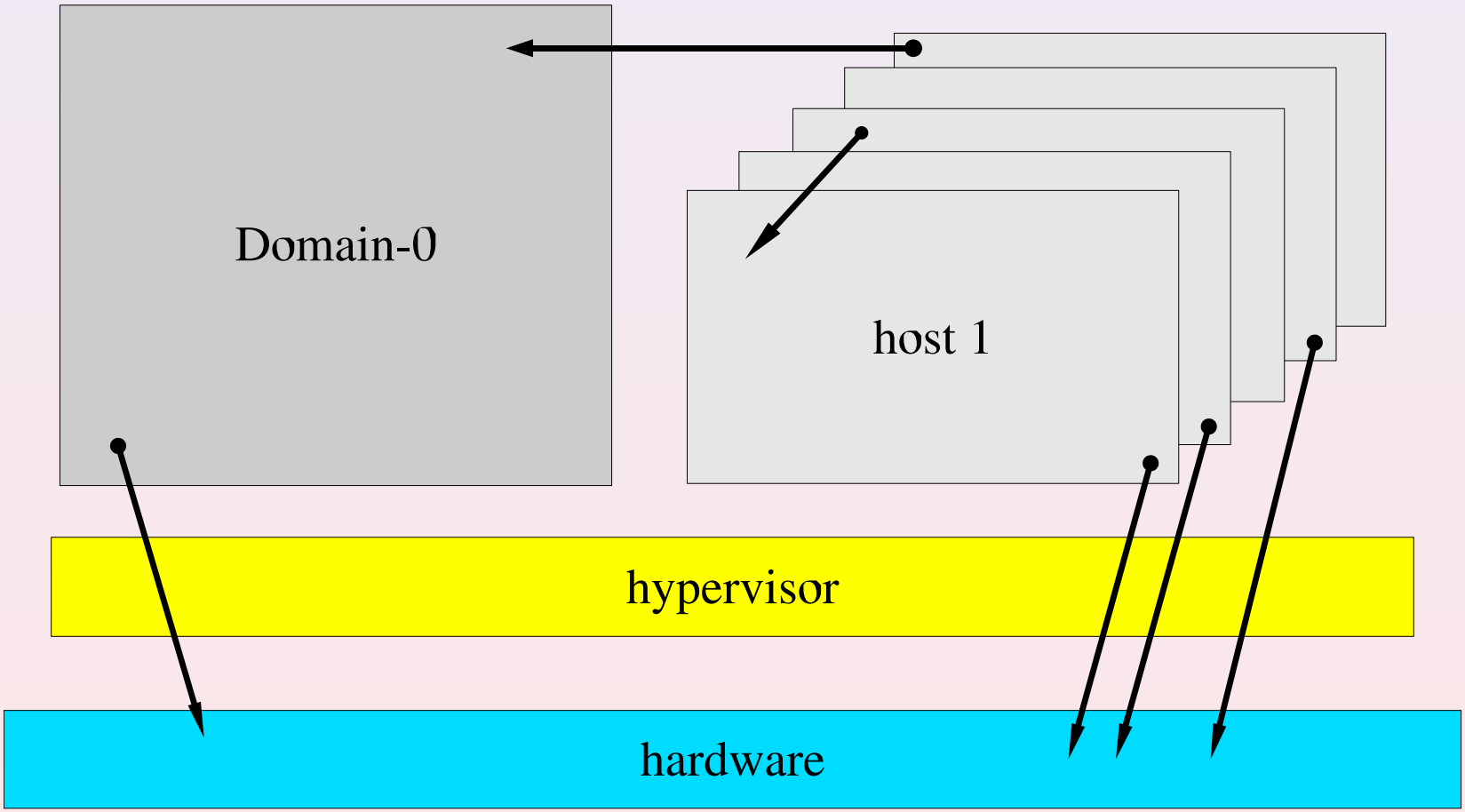
Monitor virtuálních strojů:

- velmi dynamicky se rozvíjející projekt
- stojí za ním mnohé významné firmy z IT
- je vyvíjen jako GNU

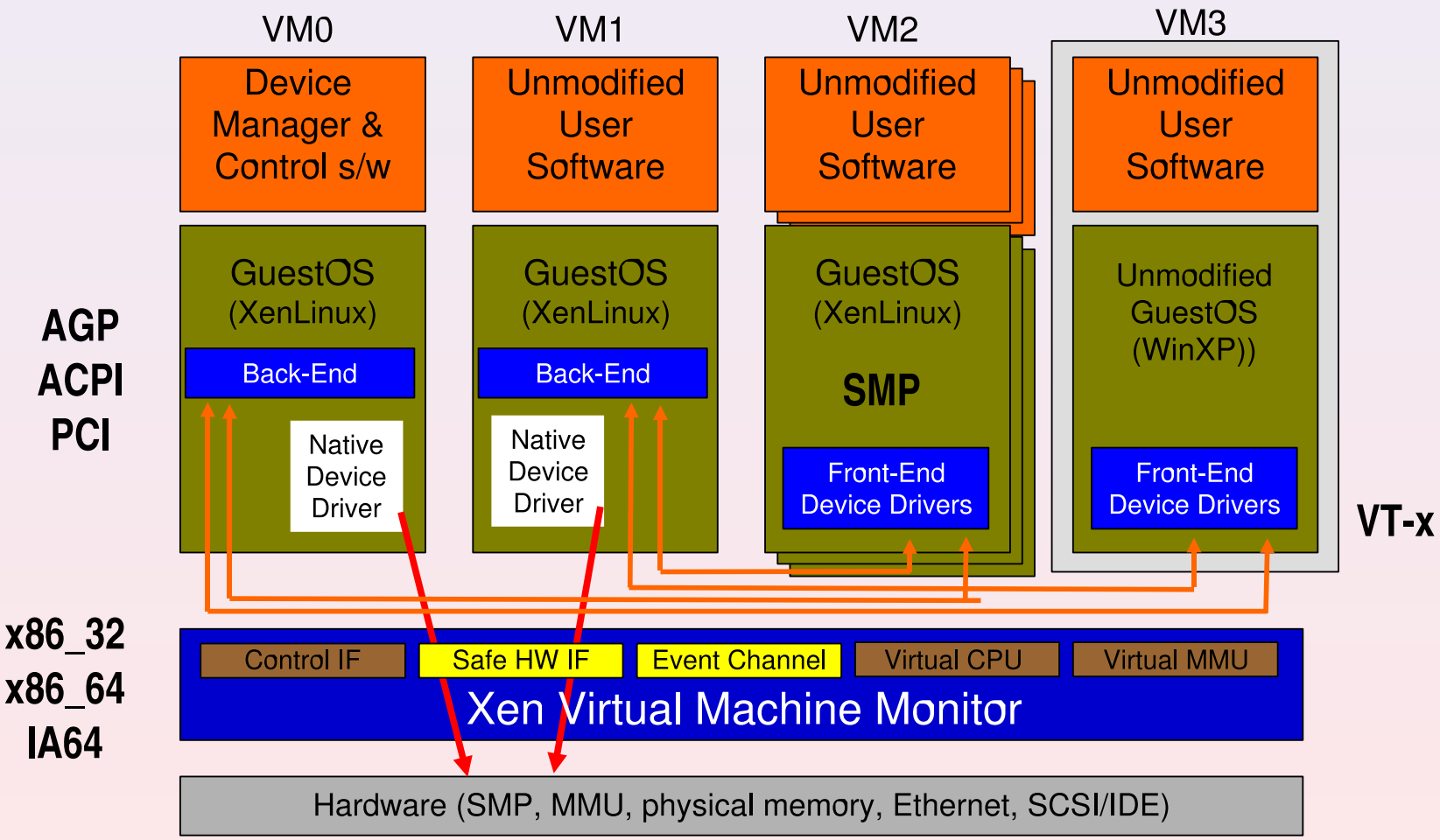
Vlastnosti Xenu

- bezpečná izolace jednotlivých virtálních strojů
- kontrola zdrojů
- upravit je potřeba jen hostovaný kernel, aplikace a knihovny zůstávají neupravené
- podpora Linux 2.4/2.6, NetBSD, FreeBSD, Plan9, Solaris
- výkonost blízká normálnímu provozu
- provozovaná architektura je x86
- živá migrace virtuálních strojů mezi nody

Zjednodušený pohled

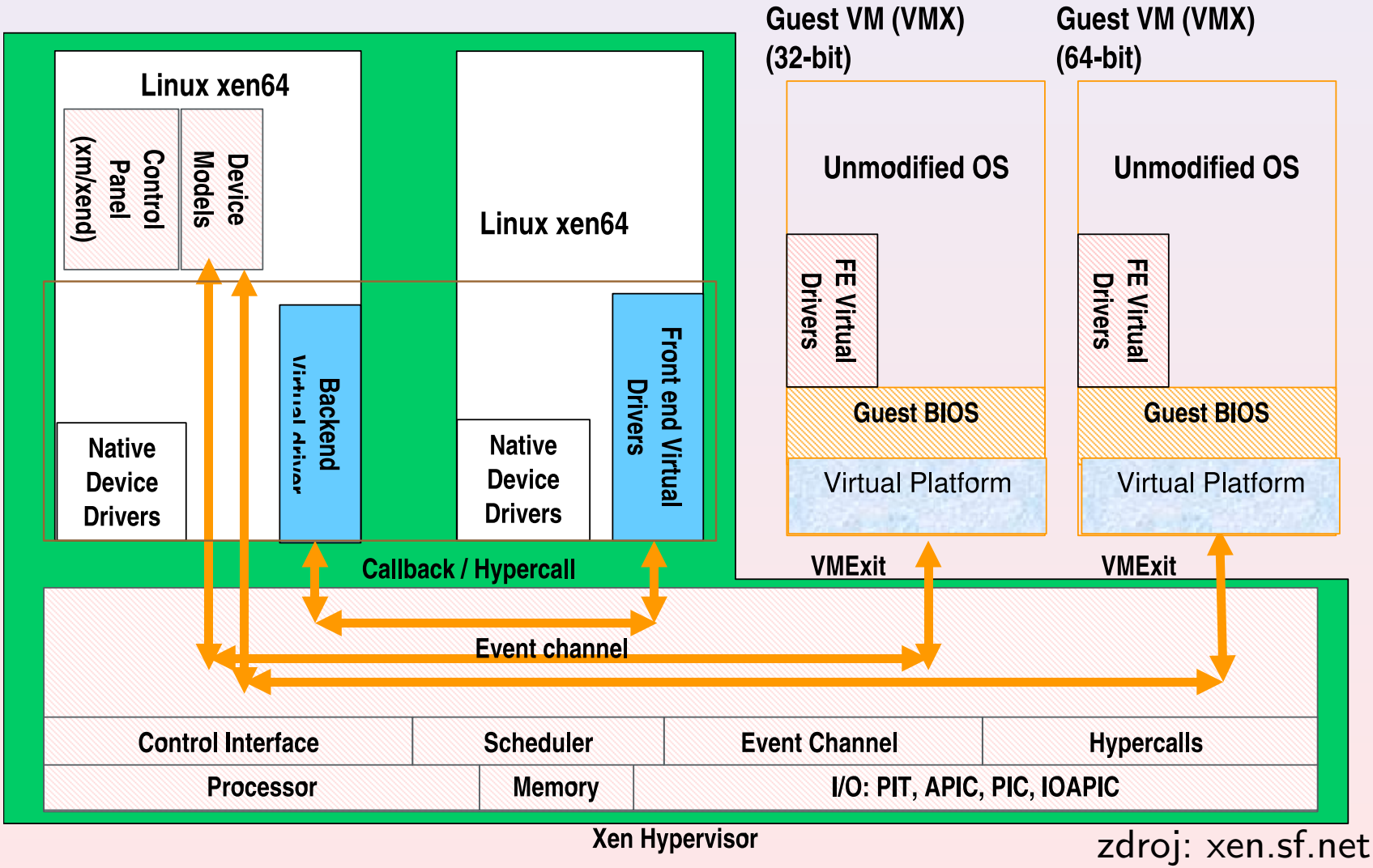


Architektura ve verzi 3.0



zdroj: xen.sf.net

Detail architektury



zdroj: xen.sf.net

Proč jsme vybrali Xen

- podpora Linuxu
- nízka režie virtualizace
- snadné ovládání
- navázání na existující prostředí
  - [illegible]
  - instalační metoda FAI
- cena

Proč jsme vybrali Xen

- podpora Linuxu
- nízka režie virtualizace
- snadné ovládání
- navázání na existující prostředí
  - souborový distribuovaný systém AFS
  - instalační metoda FAI
- cena

Proč jsme vybrali Xen

- podpora Linuxu
- nízka režie virtualizace
- snadné ovládání
- navázání na existující prostředí
  - souborový distribuovaný systém AFS
  - instalační metoda FAI
- cena

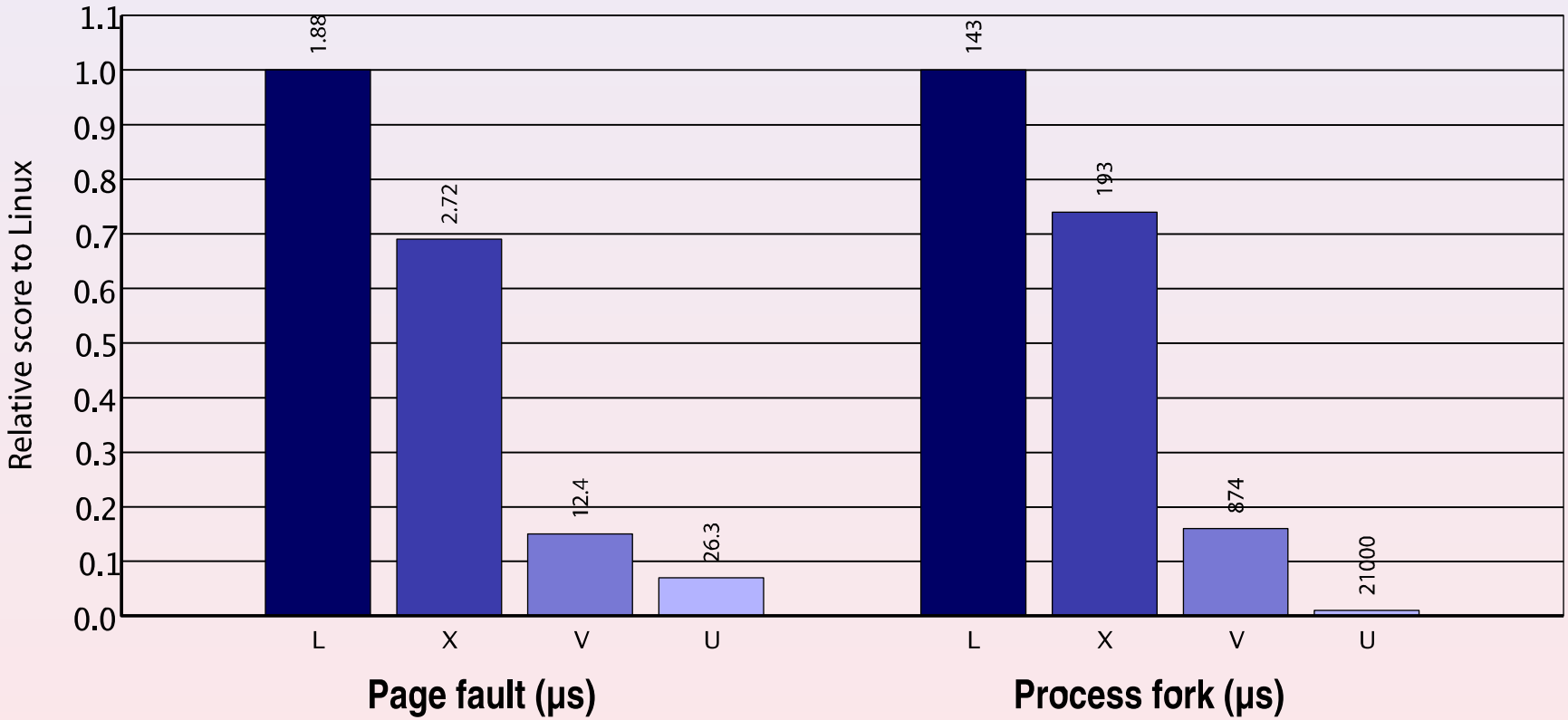
Proč jsme vybrali Xen

- podpora Linuxu
- nízka režie virtualizace
- snadné ovládání
- navázání na existující prostředí
  - souborový distribuovaný systém AFS
  - instalační metoda FAI
- cena

Proč jsme vybrali Xen

- podpora Linuxu
- nízka režie virtualizace
- snadné ovládání
- navázání na existující prostředí
  - souborový distribuovaný systém AFS
  - instalační metoda FAI
- cena

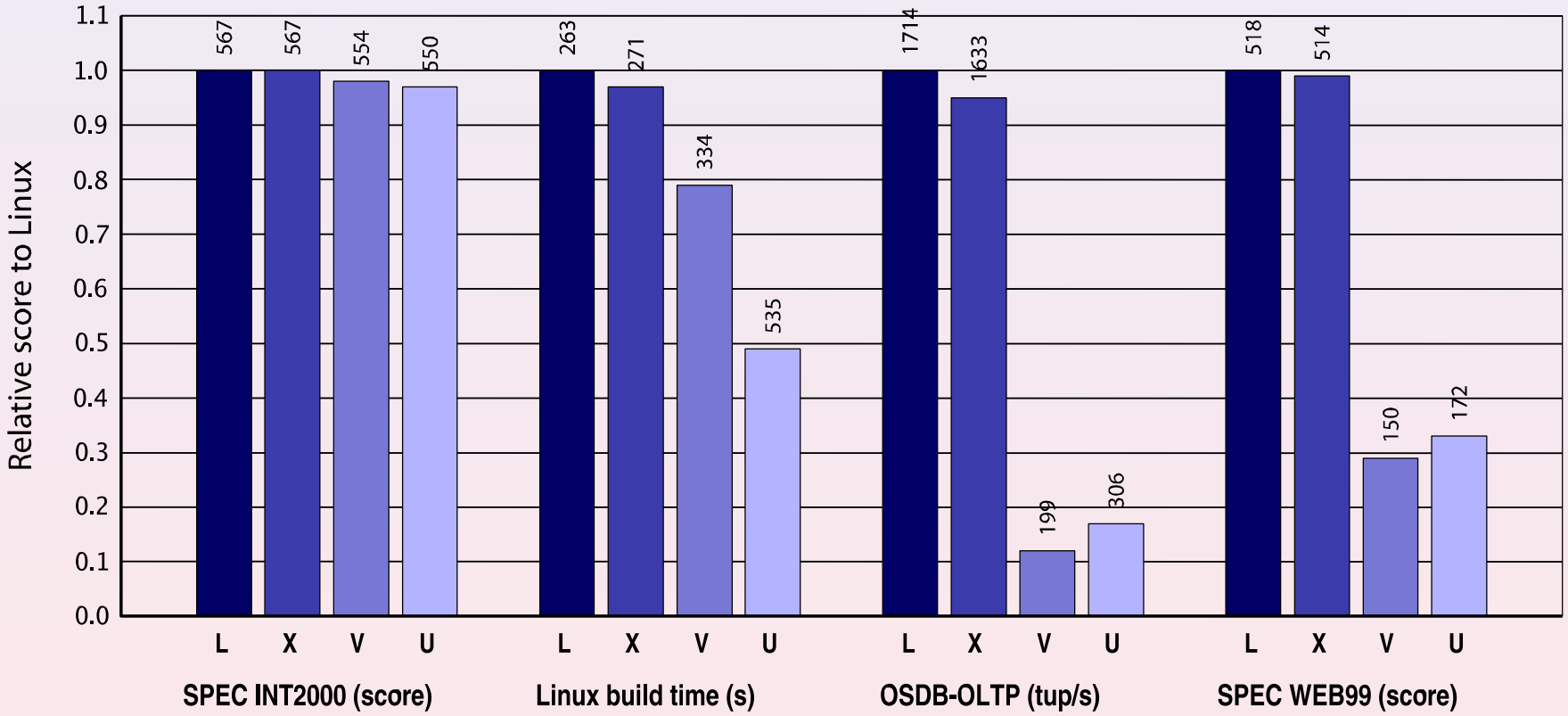
Paměť



Imbench results on Linux (L), Xen (X), VMWare Workstation (V), and UML (U)

zdroj: xen.sf.net

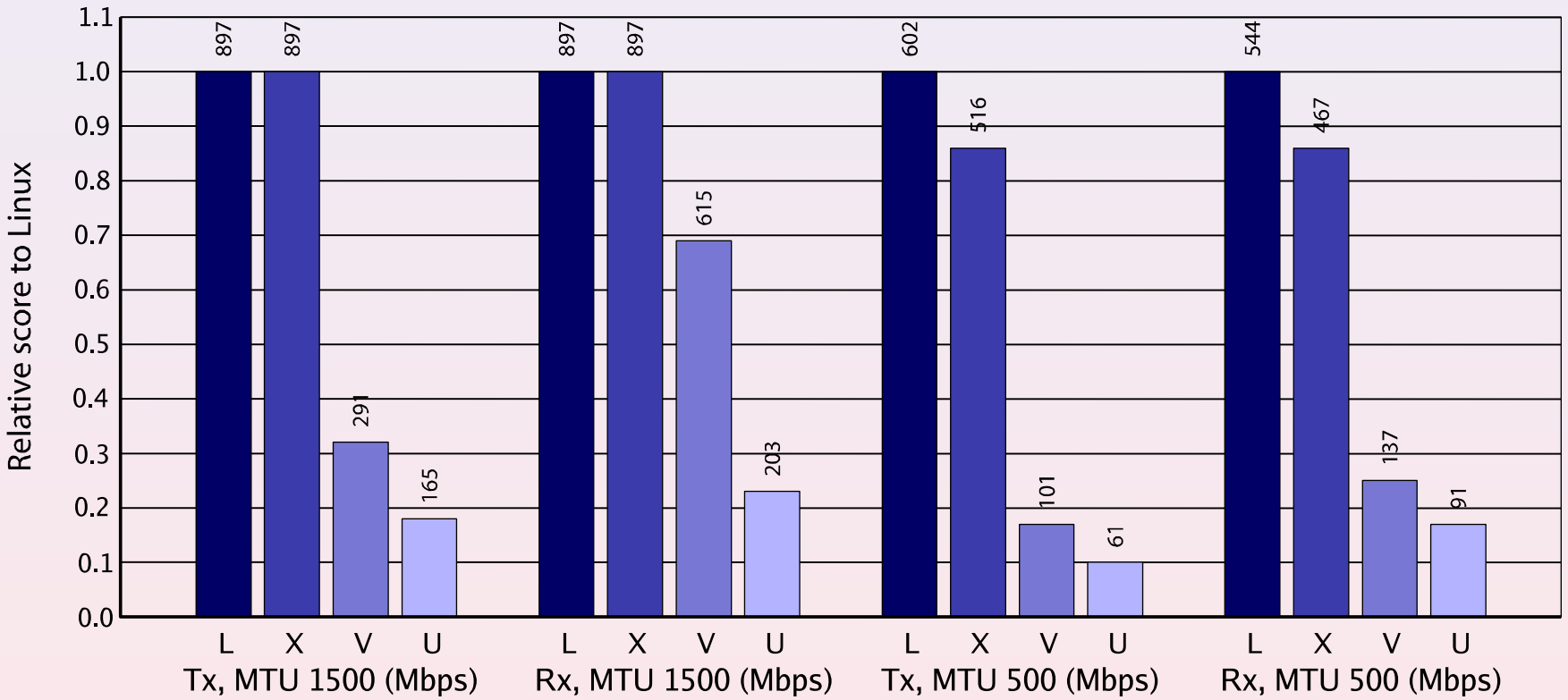
Výkonostní testy



Benchmark suite running on Linux (L), Xen (X), VMware Workstation (V), and UML (U)

zdroj: xen.sf.net

TCP testy



TCP bandwidth on Linux (L), Xen (X), VMWare Workstation (V), and UML (U)

zdroj: xen.sf.net

Náklady a zisky

- vyšší pořizovací náklady na HW
- nízké pořizovací náklady na virtuální stroj
- úspory za energie
- úspory za provoz (méně portů, zabraného místa, ...)
- zlepšení správy
- zvýšení dostupnosti
- rychlá opravitelnost

Náklady a zisky

- vyšší pořizovací náklady na HW
- nízké pořizovací náklady na virtuální stroj
- úspory za energie
- úspory za provoz (méně portů, zabraného místa, . . .)
- zlepšení správy
- zvýšení dostupnosti
- rychlá opravitelnost

Náklady a zisky

- vyšší pořizovací náklady na HW
- nízké pořizovací náklady na virtuální stroj
- úspory za energie
- úspory za provoz (méně portů, zabraného místa, . . .)
- zlepšení správy
- zvýšení dostupnosti
- rychlá opravitelnost

Náklady a zisky

- vyšší pořizovací náklady na HW
- nízké pořizovací náklady na virtuální stroj
- úspory za energie
- úspory za provoz (méně portů, zabraného místa, . . .)
- zlepšení správy
- zvýšení dostupnosti
- rychlá opravitelnost

Migrace

Migrace je přesun virtuálního stroje (hosta) mezi jednotlivými hardwarovými nody (hostitely).

Migrací virtuálních strojů získáme:

- vyšší dostupnost při údržbě
- možnost vyrovnávání zátěže nodů

Migraci rozdělujeme na:

- off-line: rychlejší, stroj je ale suspendovaný
- on-line (live): pomalejší, stroj je v provozu

Migrace

Migrace je přesun virtuálního stroje (hosta) mezi jednotlivými hardwarovými nody (hostitely).

Migrací virtuálních strojů získáme:

- vyšší dostupnost při údržbě
- možnost vyrovnávání zátěže nodů

Migraci rozdělujeme na:

- off-line: rychlejší, stroj je ale suspendovaný
- on-line (live): pomalejší, stroj je v provozu

Migrace

Migrace je přesun virtuálního stroje (hosta) mezi jednotlivými hardwarovými nody (hostitely).

Migrací virtuálních strojů získáme:

- vyšší dostupnost při údržbě
- možnost vyrovnávání zátěže nodů

Migraci rozdělujeme na:

- off-line: rychlejší, stroj je ale suspendovaný
- on-line (live): pomalejší, stroj je v provozu

Prerekvizity pro migraci

- zachování síťového segmentu
- virtuální stroj umístěn na síťovém datovém prostoru:
  - NAS: NFS, CIFS
  - SAN: Fibre Channel
  - iSCSI, síťové blokové zařízení
  - drdb

Prerekvizity pro migraci

- zachování síťového segmentu
- virtuální stroj umístěn na síťovém datovém prostoru:
  - NAS: NFS, CIFS
  - SAN: Fibre Channel
  - iSCSI, síťové blokové zařízení
  - drdb

Jak migrace probíhá

Migrace z virtuálního stroje X z nodu A na nod B:

pre-migration aktivace X na nodu A, vyběr cíle na nodu B

reservation inicializace kontejneru pro X na nodu B

pre-copy cyklické kopírování "špinavých" paměťových stránek

stop-and-copy pozastavení běhu virtuálního stroje X na nodu A, synchronizace zbylých stránek

commitment aktivace virtuálního stroje na nodu B, uvolnění virtuálního stroje na nodu A

Jak migrace probíhá

Migrace z virtuálního stroje X z nodu A na nod B:

pre-migration aktivace X na nodu A, vyběr cíle na nodu B

reservation inicializace kontejneru pro X na nodu B

pre-copy cyklické kopírování "špinavých" paměťových stránek

stop-and-copy poslední fáze kopírování:

- pozastavení virtuálního stroje X na nodu A
- přesměrování síťového provozu
- synchronizace zůstávajícího stavu

commitment aktivace virtuálního stroje na nodu B, uvolnění virtuálního stroje na nodu A

Jak migrace probíhá

Migrace z virtuálního stroje *X* z nodu *A* na nod *B*:

pre-migration aktivace X na nodu A, vyběr cíle na nodu B

reservation inicializace kontejneru pro X na nodu B

pre-copy cyklické kopírování "špinavých" paměťových stránek

stop-and-copy poslední fáze kopírování:

- pozastavení virtuálního stroje X na nodu A
- přesměrování síťového provozu
- synchronizace zůstávajícího stavu

commitment aktivace virtuálního stroje na nodu B, uvolnění virtuálního stroje na nodu A

Jak migrace probíhá

Migrace z virtuálního stroje *X* z nodu *A* na nod *B*:

pre-migration aktivace X na nodu A, vyběr cíle na nodu B

reservation inicializace kontejneru pro X na nodu B

pre-copy cyklické kopírování "špinavých" paměťových stránek

stop-and-copy poslední fáze kopírování:

- pozastaveni virtuálního stroje X na nodu A
- přesměrování síťového provozu
- synchronizace zůstávajícího stavu

commitment aktivace virtuálního stroje na nodu B, uvolnění virtuálního stroje na nodu A

Jak migrace probíhá

Migrace z virtuálního stroje *X* z nodu *A* na nod *B*:

pre-migration aktivace X na nodu A, vyběr cíle na nodu B

reservation inicializace kontejneru pro X na nodu B

pre-copy cyklické kopírování "špinavých" paměťových stránek

stop-and-copy poslední fáze kopírování:

- pozastaveni virtuálního stroje X na nodu A
- přesměrování síťového provozu
- synchronizace zůstávajícího stavu

commitment aktivace virtuálního stroje na nodu B, uvolnění virtuálního stroje na nodu A

Jak migrace probíhá

Migrace z virtuálního stroje *X* z nodu *A* na nod *B*:

pre-migration aktivace X na nodu A, vyběr cíle na nodu B

reservation inicializace kontejneru pro X na nodu B

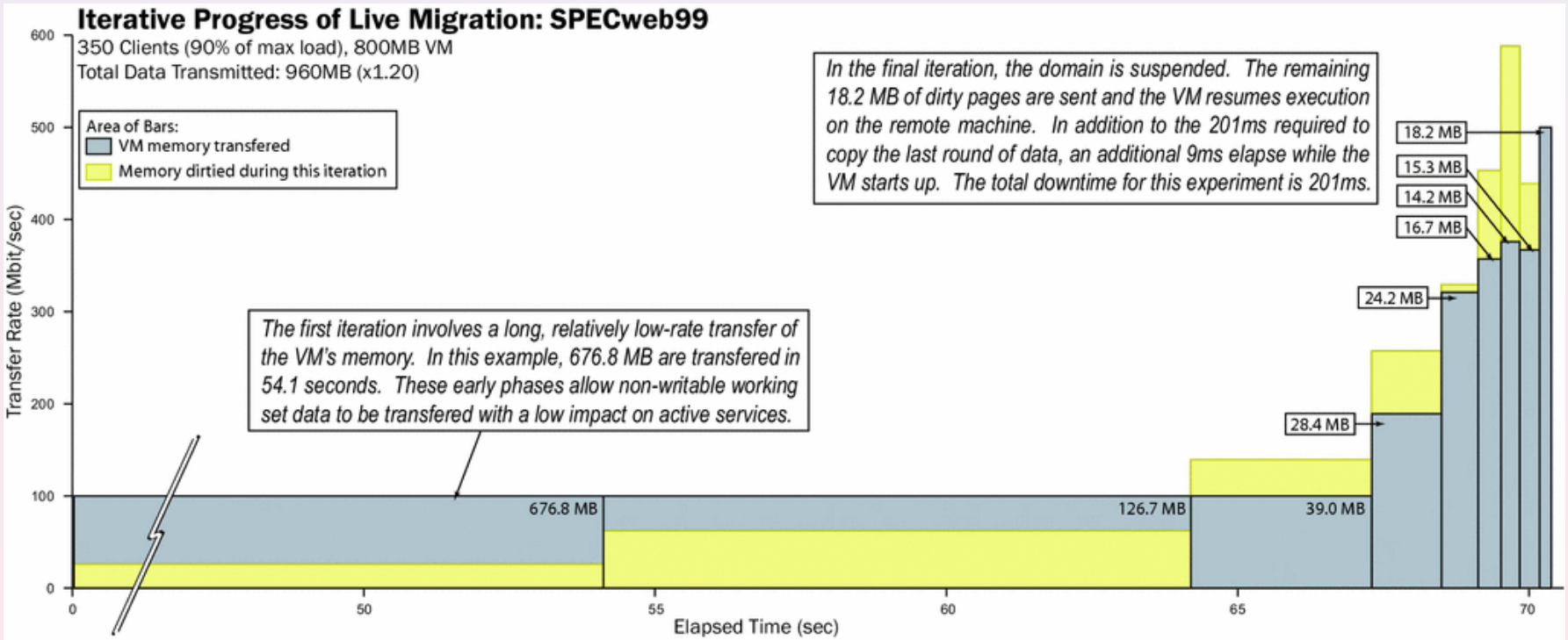
pre-copy cyklické kopírování "špinavých" paměťových stránek

stop-and-copy poslední fáze kopírování:

- pozastaveni virtuálního stroje X na nodu A
- přesměrování síťového provozu
- synchronizace zůstávajícího stavu

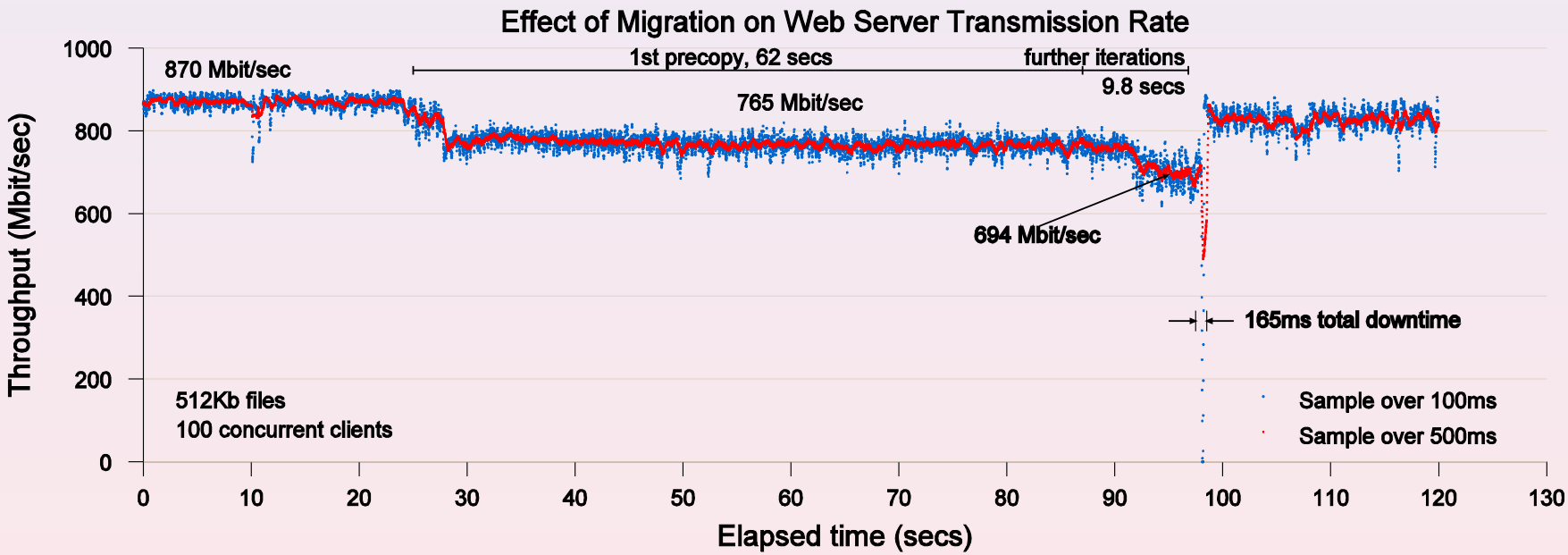
commitment aktivace virtuálního stroje na nodu B, uvolnění virtuálního stroje na nodu A

Test fáze *pre-copy*



zdroj: xen.sf.net

Migrace webového serveru



zdroj: xen.sf.net

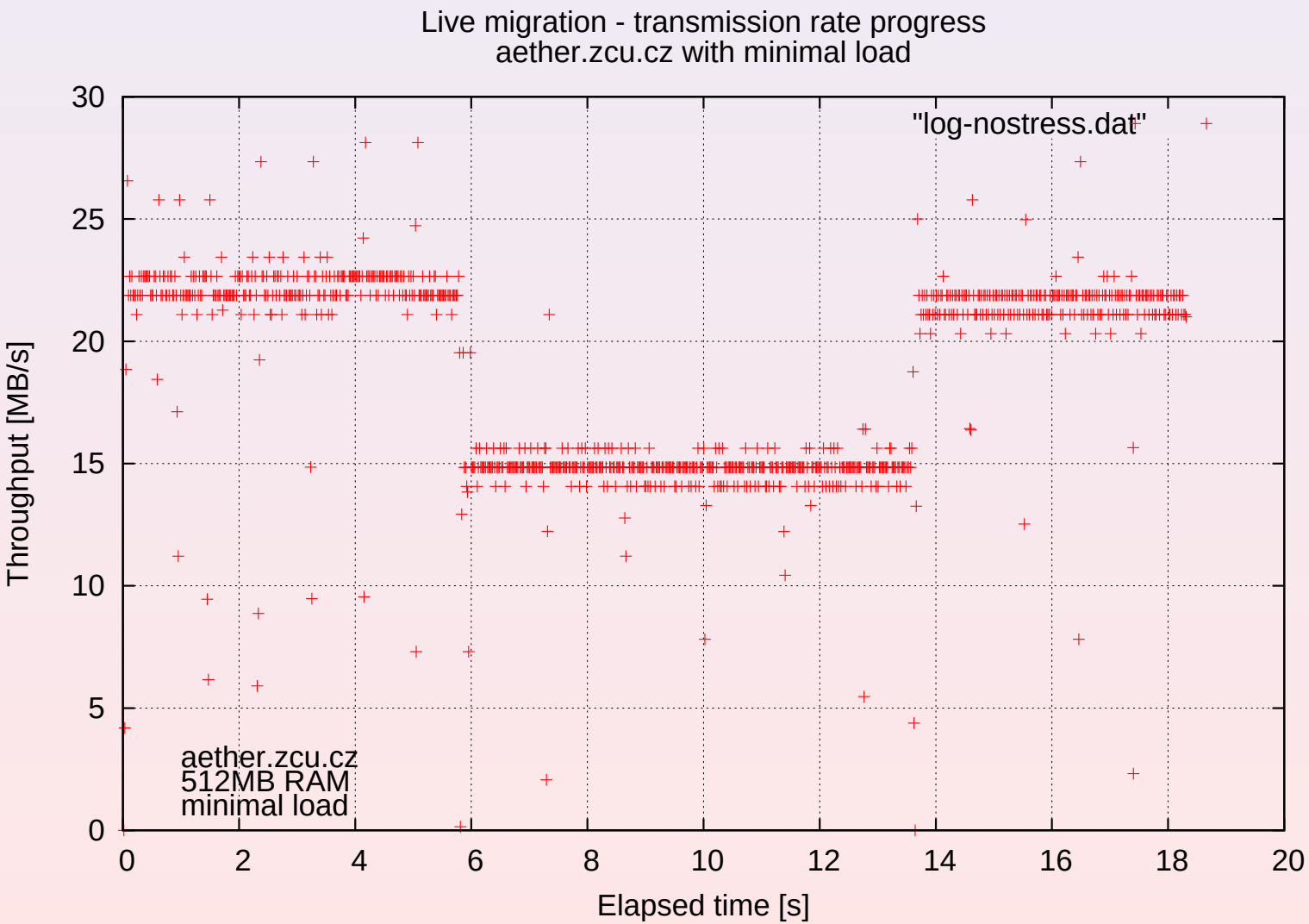
Něco z vlastní zahrádky

- Konfigurace
  - migrovaný stroj o velikosti 512MB (*aether.zcu.cz*)
  - migrace probíhala ze stroje *xen2* na stroj *xen3*
  - měřící stroj (*phoebe.zcu.cz*) umístěn na *xen3*
  - nutno nastavit forward delay: `brctl setfd br53 0`
- Měření
  - z *aether* se tahá dokola jeden soubor
  - z přírustku za 10ms se spočítá rychlost
  - použit Perl s `Time::HiRes` a `LWP::Parallel::UserAgent`
  - pro zatížení *aether* použit `stress --vm 3`

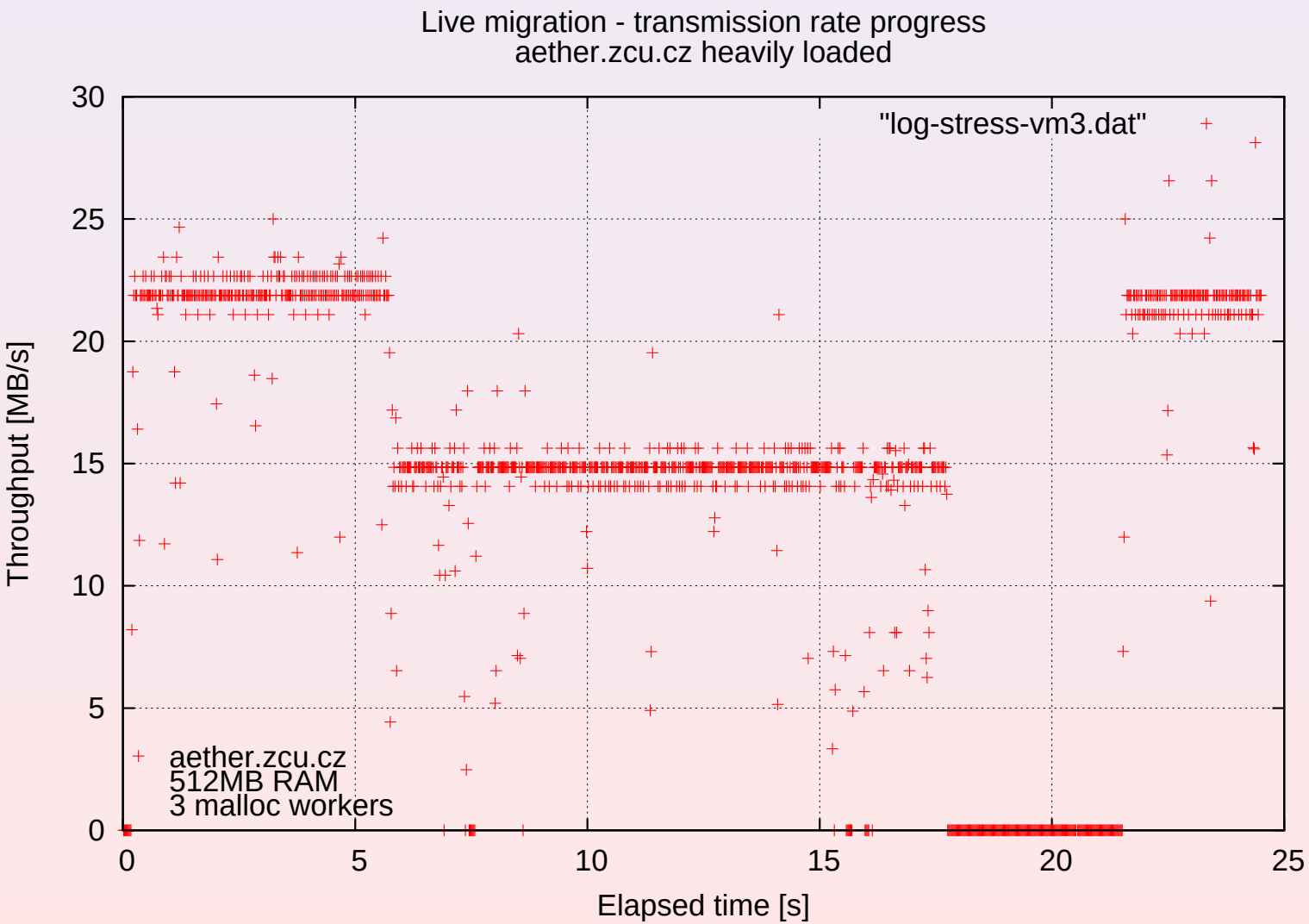
Něco z vlastní zahrádky

- Konfigurace
  - migrovaný stroj o velikosti 512MB (*aether.zcu.cz*)
  - migrace probíhala ze stroje *xen2* na stroj *xen3*
  - měřící stroj (*phoebe.zcu.cz*) umístěn na *xen3*
  - nutno nastavit forward delay: `brctl setfd br53 0`
- Měření
  - z *aether* se tahá dokola jeden soubor
  - z přírustku za 10ms se spočítá rychlost
  - použit Perl s `Time::HiRes` a `LWP::Parallel::UserAgent`
  - pro zatížení *aether* použit `stress --vm 3`

Migrace nezatížené *aether*



Migrace *aether* se zátěží



Xen na ZČU

- 3x stroj hostující Xen (verze 2.0.7)
  - 2x DualCore Xeon na 3GHz
  - 4GB RAM, 2x Gbit ethernet, 2x 80GB SATA v SW RAIDu
  - FibreChannel karta
- celkem 26 virtálních strojů (8 v ostrém provozu)
  - 64-800MB RAM
  - 10-20GB HDD (včetně 2GB swap souboru) na FC
  - 1x virtuální CPU, 1x virtuální ethernet připojen do VLAN
- správa virtuálních strojů
  - instalace FAI
  - konfigurace na NFS
- do budoucna
  - přechod na řadu 3.x
  - vyzkoušet možnosti technologie IntelVT nebo Pacifica
  - vyrovnávání zátěže mezi nody

Xen na ZČU

- 3x stroj hostující Xen (verze 2.0.7)
  - 2x DualCore Xeon na 3GHz
  - 4GB RAM, 2x Gbit ethernet, 2x 80GB SATA v SW RAIDu
  - FibreChannel karta
- celkem 26 virtálních strojů (8 v ostrém provozu)
  - 64-800MB RAM
  - 10-20GB HDD (včetně 2GB swap souboru) na FC
  - 1x virtualní CPU, 1x virtualní ethernet připojen do VLAN
- správa virtuálních strojů
  - instalace FAI
  - konfigurace na NFS
- do budoucna
  - přechod na řadu 3.x
  - vyzkoušet možnosti technologie IntelVT nebo Pacifica
  - vyrovnávání zátěže mezi nody

Xen na ZČU

- 3x stroj hostující Xen (verze 2.0.7)
  - 2x DualCore Xeon na 3GHz
  - 4GB RAM, 2x Gbit ethernet, 2x 80GB SATA v SW RAIDu
  - FibreChannel karta
- celkem 26 virtálních strojů (8 v ostrém provozu)
  - 64-800MB RAM
  - 10-20GB HDD (včetně 2GB swap souboru) na FC
  - 1x virtualní CPU, 1x virtualní ethernet připojen do VLAN
- správa virtuálních strojů
  - instalace FAI
  - konfigurace na AFS
- do budoucna
  - přechod na řadu 3.x
  - vyzkoušet možnosti technologie IntelVT nebo Pacifica
  - vyrovnávání zátěže mezi nody

Xen na ZČU

- 3x stroj hostující Xen (verze 2.0.7)
  - 2x DualCore Xeon na 3GHz
  - 4GB RAM, 2x Gbit ethernet, 2x 80GB SATA v SW RAIDu
  - FibreChannel karta
- celkem 26 virtálních strojů (8 v ostrém provozu)
  - 64-800MB RAM
  - 10-20GB HDD (včetně 2GB swap souboru) na FC
  - 1x virtualní CPU, 1x virtualní ethernet připojen do VLAN
- správa virtuálních strojů
  - instalace FAI
  - konfigurace na AFS
- do budoucna
  - přechod na řadu 3.x
  - vyzkoušet možnosti technologie IntelVT nebo Pacifica
  - vyrovnávání zátěže mezi nody

Xen na ZČU

- 3x stroj hostující Xen (verze 2.0.7)
  - 2x DualCore Xeon na 3GHz
  - 4GB RAM, 2x Gbit ethernet, 2x 80GB SATA v SW RAIDu
  - FibreChannel karta
- celkem 26 virtálních strojů (8 v ostrém provozu)
  - 64-800MB RAM
  - 10-20GB HDD (včetně 2GB swap souboru) na FC
  - 1x virtualní CPU, 1x virtualní ethernet připojen do VLAN
- správa virtuálních strojů
  - instalace FAI
  - konfigurace na AFS
- do budoucna
  - přechod na řadu 3.x
  - vyzkoušet možnosti technologie IntelVT nebo Pacifica
  - vyrovnávání zátěže mezi nody

Xen na ZČU

- 3x stroj hostující Xen (verze 2.0.7)
  - 2x DualCore Xeon na 3GHz
  - 4GB RAM, 2x Gbit ethernet, 2x 80GB SATA v SW RAIDu
  - FibreChannel karta
- celkem 26 virtálních strojů (8 v ostrém provozu)
  - 64-800MB RAM
  - 10-20GB HDD (včetně 2GB swap souboru) na FC
  - 1x virtualní CPU, 1x virtualní ethernet připojen do VLAN
- správa virtuálních strojů
  - instalace FAI
  - konfigurace na AFS
- do budoucna
  - přechod na řadu 3.x
  - vyzkoušet možnosti technologie IntelVT nebo Pacifica
  - vyrovnávání zátěže mezi nody

Xen na ZČU

- 3x stroj hostující Xen (verze 2.0.7)
  - 2x DualCore Xeon na 3GHz
  - 4GB RAM, 2x Gbit ethernet, 2x 80GB SATA v SW RAIDu
  - FibreChannel karta
- celkem 26 virtálních strojů (8 v ostrém provozu)
  - 64-800MB RAM
  - 10-20GB HDD (včetně 2GB swap souboru) na FC
  - 1x virtualní CPU, 1x virtualní ethernet připojen do VLAN
- správa virtuálních strojů
  - instalace FAI
  - konfigurace na AFS
- do budoucna
  - přechod na řadu 3.x
  - vyzkoušet možnosti technologie IntelVT nebo Pacifica
  - vyrovnávání zátěže mezi nody

Xen na ZČU

- 3x stroj hostující Xen (verze 2.0.7)
  - 2x DualCore Xeon na 3GHz
  - 4GB RAM, 2x Gbit ethernet, 2x 80GB SATA v SW RAIDu
  - FibreChannel karta
- celkem 26 virtálních strojů (8 v ostrém provozu)
  - 64-800MB RAM
  - 10-20GB HDD (včetně 2GB swap souboru) na FC
  - 1x virtualní CPU, 1x virtualní ethernet připojen do VLAN
- správa virtuálních strojů
  - instalace FAI
  - konfigurace na AFS
- do budoucna
  - přechod na řadu 3.x
  - vyzkoušet možnosti technologie IntelVT nebo Pacifica
  - vyrovnávání zátěže mezi nody

Krátká diskuse během technické pauzy

aneb reboot do života Xenu.