Rozsáhlé distribuované systémy v reálném světě

Luděk Matyska

Kolokviální profesorská přednáška

Fakulta informatiky, Masarykova univerzita

Motto

- Much of the present interest in parallel processing stems from a now well–documented technological trend—the declining cost of processing, storage, and communications.

Motto

- Much of the present interest in parallel processing stems from a now well–documented technological trend—the declining cost of processing, storage, and communications.

- W. A. Kornfeld, C. E. Hewitt, IEEE Transactions on Systems, Man, and Cybernetics, 11(1), January 1981

Internet jako distribuovaný systém

- Vlastní přenosový systém
  - Trasy a aktivní prvky
  - Vysoká míra homogenity
  - Malý počet jednoduchých služeb
- Zapojení koncových stanic
  - Heterogenita
  - Prakticky nekonečné množství služeb
  - Přímé zapojení koncových uživatelů

Vlastnosti

- Postupné stírání rozdílu mezi úzce propojenými – paralelními – a volně propojenými – distribuovanými – systémy
  - Významný vliv kvality sítě
    - Relativně vzhledem k parametrům paralelních (super)počítačů
- >100 milionů propojených uzlů
  - Spíše teoretická veličina
  - Peer to peer sítě i jednotky milionů
  - Silně koordinované systémy podstatně menší

Naše motivace

- Postavit a provozovat rozsáhlou výpočetní a úložnou infrastrukturu
- Primární přístup prostřednictvím úloh
  - Tvořeny jedním či více procesy
  - Pracují s nějakými daty
  - Dávkové nebo interaktivní
- Pokročilé použití: virtuální clustery
  - Manipulace, zpřístupnění, …

Rozsáhlá infrastruktura – Grid

- Stovky propojených míst
- Desítky až stovky tisíc zapojených uzlů
- Koordinované využití
  - Koordinace na technologické i organizační úrovni
  - Různorodé požadavky na bezpečnost dat i zabezpečení infrastruktury
- Výrazně vyšší míra kontroly než peer to peer sítě
  - Cenou ztráta flexibility a části autonomie uzlů
  - Dependable computing: je možné se spolehnout

Funkční požadavky na prostředí

- Jednoznačně identifikovat úlohu
  - A to i za předpokladu více nezávislých vstupů úloh do infrastruktury
- Naplánovat úlohu
  - Nezbytná znalost stavu infrastruktury
  - Data a jejich rozložení
- Sledovat chování úlohy
  - Průchod komponentami
- Správa dat
  - Ukládání, transport, replikace
- Zajistit bezpečnost celého systému
  - Zdroje, procesy i data

Globální identifikace úlohy

- Není možná centralizovaná služba
  - Výkonnost i spolehlivost
- Statisticky jedinečný identifikátor
  - URL místa vstupu úlohy plus hash dalších parametrů včetně času
    - https://tigerman.cnaf.infn.it:9000/ec8dEpe696avU557gvHWg
  - Monitorování na uvedeném URL
  - Aplikováno v projektech EU EGEE

Naplánování úlohy

- Centrální (meta) plánovač
  - Předpokládá kompletní znalost stavu infrastruktury
  - Úzké místo
    - Škálovatelnost: milion úloh denně znamená 12 úloh za sekundu
- Kooperující plánovače
  - Latence rozhodování
- Soupeřící plánovače

  Fibich, Matyska, Rudová: Model of Grid Scheduling Problem, Exploring Planning and Scheduling for Web Services, Grids and Autonomic Computing, AAAI Press, 2005

Sledování stavu infrastruktury

- Nezbytný předpoklad pro plánovač
- Senzory nasazené na jednotlivé komponenty
  - Infrastruktura sběru a agregace informací
  - Capability Based Grid Monitoring Architecture (C-GMA)

    A. Ceccanti et al: Towards Scalable and Interoperable Grid Monitoring Infrastructure, 1st CoreGID Integration Workshop, 2005

    Sitera et al: Capability and Attribute Based Grid Monitoring Architecture. Proc. CGW04, 2005
- Aplikačně orientované sledování

  L. Matyska et al: Non-centralized Grid Infrastructure Monitoring, Dagstuhl Seminar, 2004

  Holub et al: Grid infrastructure monitoring as reliable information service. LNCS 3165, 2004

Sledování úloh

- Průchod řadou komponent
  - Pouze lokální informace
- Služba Logging and Bookkeeping
  - Události plynou od komponent
  - Stavový automat s korekcí chyb

    G. Avellino et al: The DataGrid Workload Management Systém: Challenges and Results, J. Grid Computing 2(4) 2005

    Kouřil et al: Distributed Tracking, Storage, and Re-use of Job State Information on the Grid, CHEP 2004
- Job Provenance
  - Trvalé úložiště informací o stavu úloh

    Dvořák et al: gLite Job Provenance. IPAW 2006, LNCS 4145, 2006

Bezpečnost

- Identifikace uživatele
  - Používá zdroje, které nikdy neviděl, poskytované institucemi, s nimiž nemá žádný bezprostřední formální vztah

    Kouřil, Procházka, Matyska: Experience with PKI in a Large-scale Distributed Environment, EUNIS07, 2007
- Oprávnění přístupu
  - Granularita – příliš mnoho uživatelů i komponent (služeb)

Utkvělé představy o vlastnostech

- Komponenty distribuovaného systému jsou spolehlivé
- Topologie sítě/distribuovaného systému se nemění
- Síť/Distribuovaný systém je homogenní
- Distribuovaný systém má jednoho správce
- Cena přenosu dat je nulová
- Zpoždění je nulové
- Propustnost je nekonečná
- Síť je bezpečná
- Čas na prvcích je synchronizován

Sledování stavu infrastruktury

- Senzory na komponentách
  - Spolehlivost senzorů
  - Potřebujeme sběrnou/agregační infrastrukturu
    - Kdo hlídá hlídače?
  - C-GMA pro agregaci různorodých infrastruktur
- Aplikačně orientované sledování
  - Místo senzoru „malá“ aplikace
    - Ověří všechny komponenty systému
    - Výsledkem data o stavu
    - Sbírána a vyhodnocena nezávisle na „aplikaci“
- Nestačí jeden systém, nutná křížová kontrola

Plánování úloh – problémy

- Rychlost přijetí úloh systémem
  - 5000 úloh trvá 7 minut při propustnosti milion úloh denně
- Rychlost zpracování plánovačem
  - Vylučuje jednoduché nasazení složitého plánování
    - Triviální plánovače versus nepoužitelné
- Neaktuálnost informací o infrastruktuře
  - Neexistence globálního stavu (rychlost světla)
  - Chyby senzorů i konfigurací

    „Černé díry“

Plánování úloh – MatchMaking

- Párování požadavků úloh (např. paměť, počet procesorů, …) a parametrů zdrojů
  - Zpravidla „First fit“
- Sémantika parametrů
  - Ad hoc
  - Globální dohoda – Glue Schema
- Velmi závislé na kvalitě informací o úlohách i zdrojích
- V podstatě velmi primitivní plánovač
  - Žádné pokročilé schopnosti, např. dodržení požadavku na čas dokončení úlohy

Pilotní úlohy

- Snaha odstranit nedostatky spolehlivosti
- Plánuje se „obálka“, tj. prázdná úloha
  - Pokud se „obálka“ skutečně spustí (tj. nezasekne se ani se neztratí), pak si ověří prostředí, v němž byla spuštěna a „šáhne“ si na skutečnou úlohu
- V podstatě implementace konceptu aplikačně-orientovaného sledování infrastruktury
  - Obálka úlohy je sama testem
  - Ztracené obálky se nepočítají
- Umožňuje implementovat velmi důmyslné algoritmy plánování na straně uživatele
  - Ovšem v realitě se tato možnost zatím nevyužívá

Událostmi řízené plánovaní

- Různé typy událostí:
  - Příchod nové úlohy
  - Úspěšné ukončení úlohy
  - Zhroucení úlohy či uzlu
  - Ztráta/zmizení úlohy
- Událost způsobí pře-plánování
  - Stavíme skutečný „plán“ úloh
- Možnost průběžné optimalizace plánu
  - Běží mezi příchodem událostí
  - Rychlá reakce po příchodu události
- Umí pracovat i v nekooperativním módu

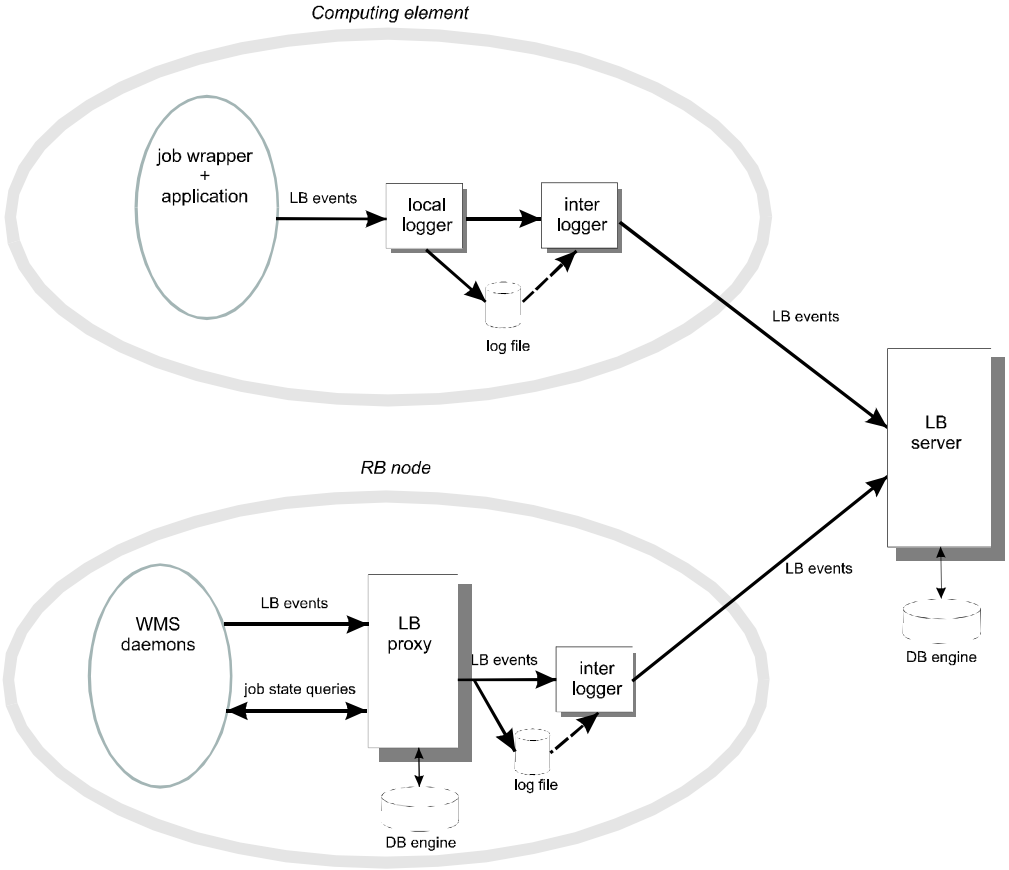
Klusáček, Rudová, Matyska: Local Search for Deadline Driven Grid Scheduling, MEMICS 2007

Klusáček, Matyska, Rudová: Alea—Grid Scheduling Simulation Enivronment, LNCS 4967, 2007

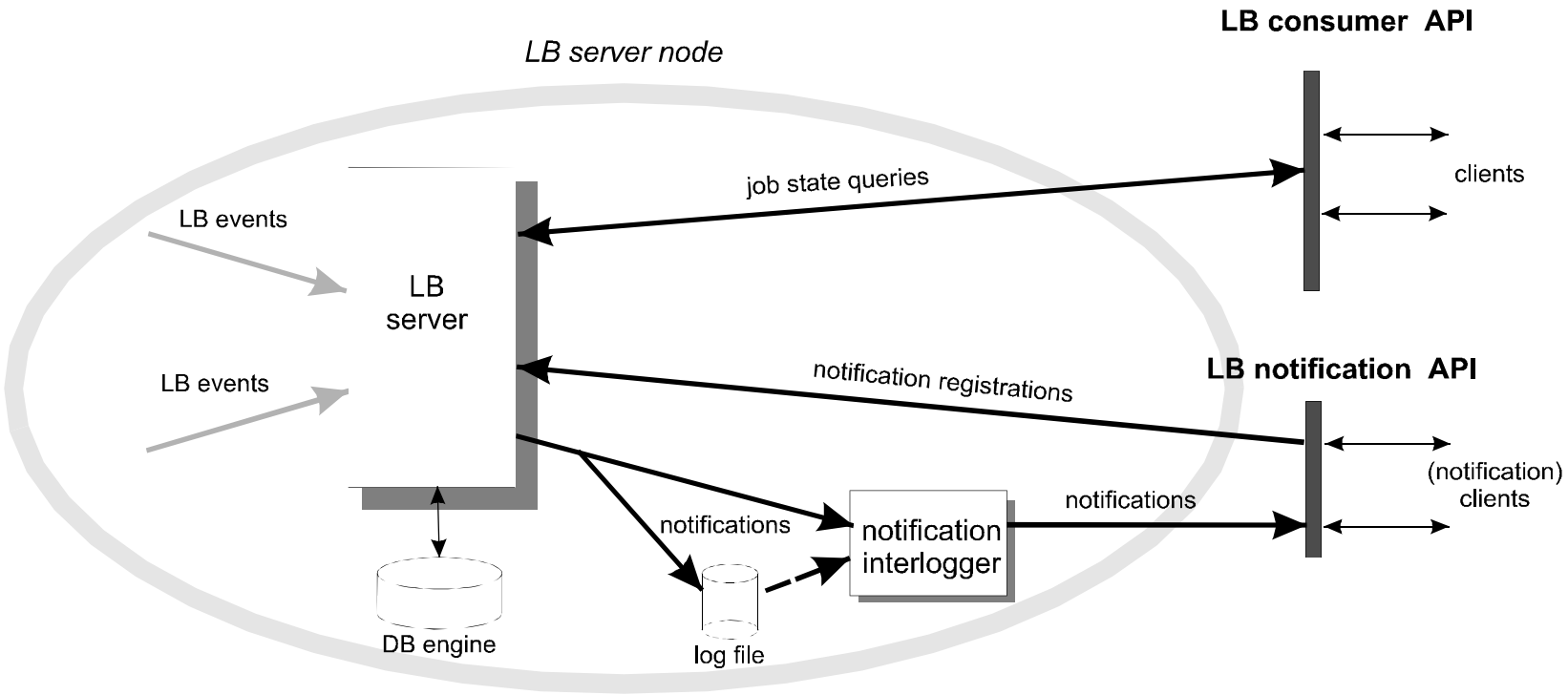
Logging and Bookkeeping (LB)

- Koncept a služba vyvinutá v rámci projektů EU
- Pracuje rovněž s konceptem *událostí*
  - Lokální změny stavu úlohy
  - Předání úlohy mezi komponentami
- Události sbírá centralizovaná komponenta
  - Bookkeeping server
- Počítá s nespolehlivostí distribuovaného prostředí
  - Nesynchronizovaný čas
  - Události mohou přijít v libovolném pořadí
  - Události se mohou ztratit (i přes zabezpečení komunikační protokoly)

Sběr dat


Computing element
job wrapper
+
application
LB events
local
logger
inter
logger
log file
LB events
LB
server
RB node
WMS
daemons
LB events
job state queries
LB
proxy
LB events
inter
logger
log file
DB engine
LB events
DB engine

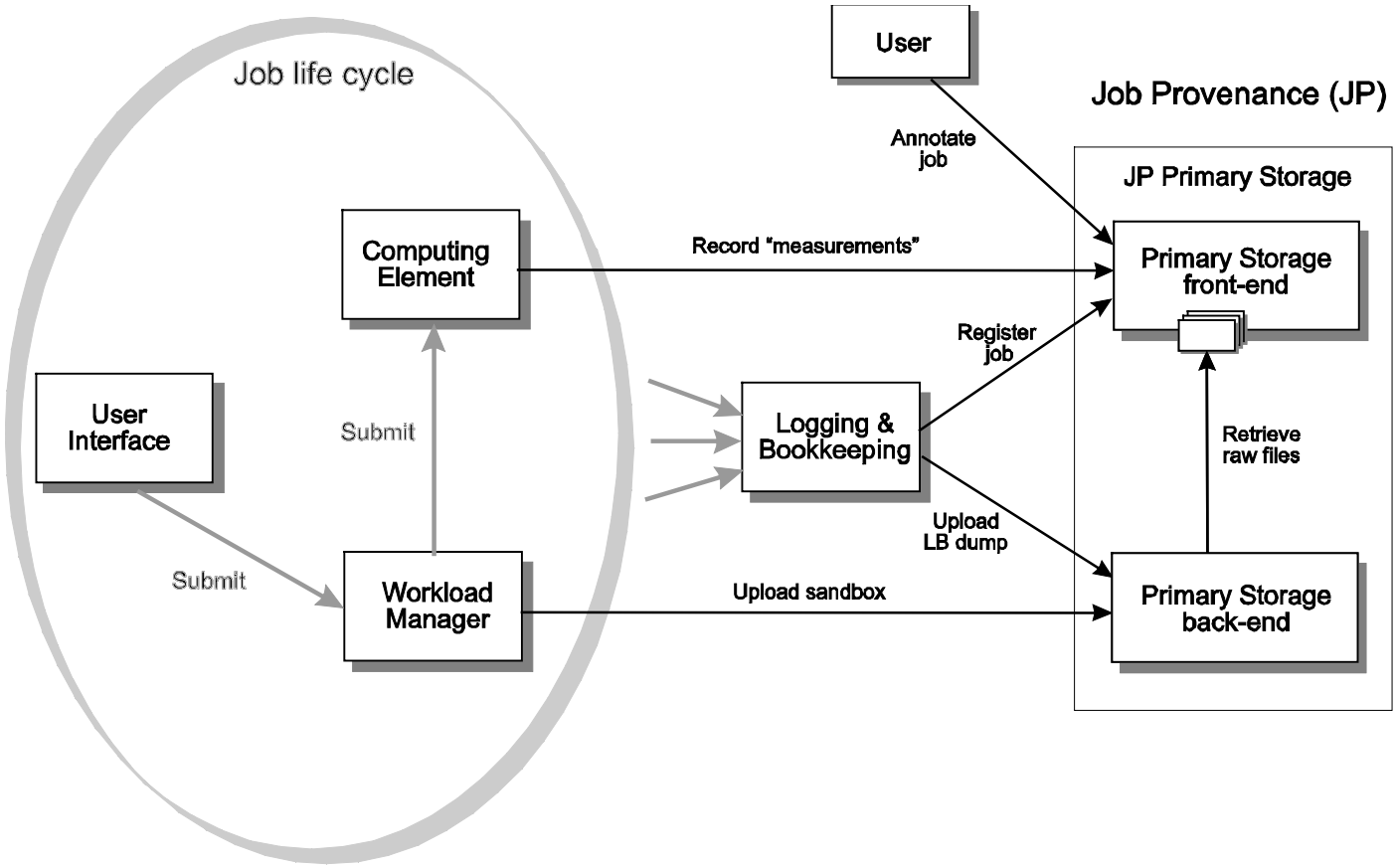
Zpřístupnění dat



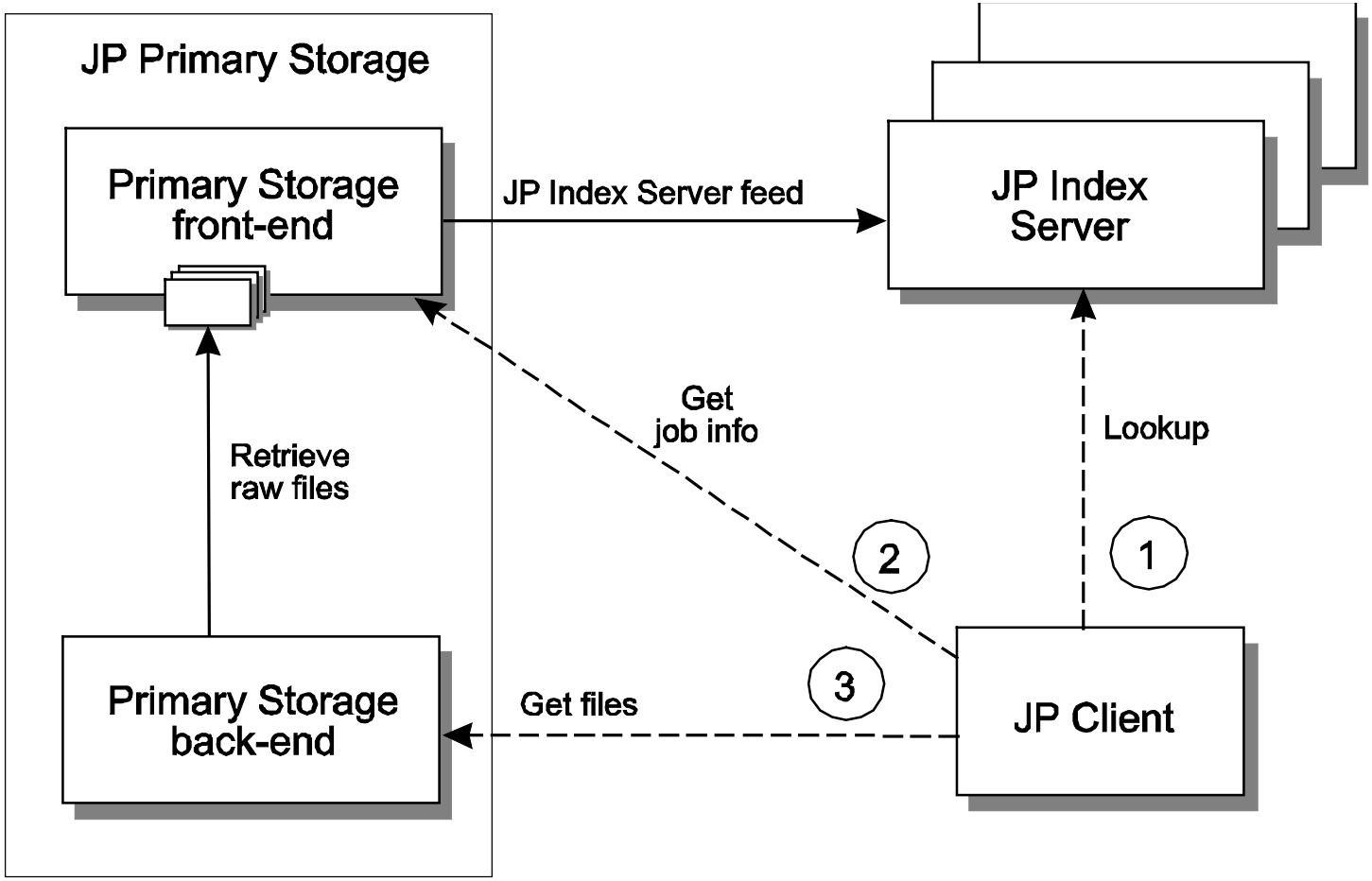
Job Provenance

- LB drží data jen pokud je úloha aktivní
  - JP reaguje na potřeba dlouhodobého ukládání informací o skončených úlohách
- Dvě komponenty
  - JP Primary Server – optimalizován pro ukládání
    - Rozšířená kopie záznamuz LB
  - JP Index Server – optimalizován pro dotazy
    - Index k jedné či více JPPS
- Zpřístupnění dat s měnícími se formáty
  - Koncept plug-inů, zpracovávají přicházející data
  - Umožňuje rozšiřování sémantiky ukládaných dat
- Uživatelské anotace
  - Možné přidávat i po skončení úlohy

JP Primary Server



JP Index Server



Nasazení LB a JP

- LB v produkčním provozu na desítkách míst po Evropě i mimo ni
- JP v experimentálním nasazení
  - Účast na *Provenance Challenges*
    - V rámci EU projektu Provenances
  - Integrace se systémy zpřístupnění a manipulace s uloženými údaji
    - Charon (ČR)
    - DashBoard (EU)

Bezpečnost – identifikace uživatele

- Centralizovaná řešení (Kerberos)
  - Špatná škálovatelnost
  - Složité přes několik administrativních domén
- Distribuovaná řešení
  - Systém veřejných klíčů, Public Key Infrastructure (PKI)
  - Role Certifikačních Autorit (CA)
    - Organizace, která potvrdí spojení konkrétní osoby a jejího veřejného klíče
    - Udržuje seznam již neplatných (revokovaných) klíčů
  - Akceptujeme-li CA, nepotřebujeme sami identifikovat uživatele
    - Musí existovat způsob certifikace CA – EUGridPMA, IGTF

Problémy PKI

- Poměrně složité získání certifikátu
  - Jednoduchost by ale znamenala riziko ztráty důvěryhodnosti
- Ochrana soukromého klíče
  - Binární data, uložená na disku – riziko zkopírování
    - Ochrana heslem
  - Uložení na hardwarovém tokenu (chipová karta, USB klíč)
  - Klíč nutný k identifikaci osoby, která zadala úlohu
    - Ovšem úloha může velmi dlouho čekat ve frontě, jak zajistit bezpečnost v tomto případě
- Proxy certifikáty
  - Vygenerované krátkodobé klíče, potvrzené majitelem dlouhodobého klíče
  - Problém obnovení proxy certifikátu

Tokeny

- Nemusí dlouhodobý klíč vůbec zpřístupnit
  - Přímo provádějí nezbytné operace
- Praktické problémy
  - Koncept původně vyvinut pro webové služby, kde stačila jednorázová identifikace
    - Přístup k Internetovému bankovnictví
    - Nevhodné pro častou opakovanou autentizaci
      - Např. s každou úlohou
  - Proxy certifikát nepomůže – jak automaticky obnovovat?
- Důsledek: USB klíč trvale zapojen do stolního počítače s přístupem přes síť
  - Kompletní popření výhod tokenů

Zkušenosti z národního projektu poskytování tokenů uživatelům

Federace

- Místo jednotného autentizačního mechanismu jejich kooperace
  - Využití spolehlivých autentizačních mechanismů domovských institucí uživatelů
  - Při autentizaci uživatele vůči konkrétní službě se požadavek přesměruje na domovskou organizaci – *Identity Provider (IdP)*
    - Ta ověří a vydá verdikt
    - Součástí odpovědi mohou být i atributy uživatele
      - Částečná anonymizace, IdP sdělí např. zda se jedná o studenta či učitele, ale nesdělí totožnost
        - » Identitu lze dohledat
- Navázání na PKI – on-line certifikační autorita
  - Stále problém s obnovením krátkodobého certifikátu

Procházka et al: Transaprent Security for Collaborative Envirnments. 3rd Collaborative Computing, 2007

Složité úlohy plánování

- Advanced reservation
  - Potřebujeme úlohu spustit v určitém čase v budoucnosti
  - Nezbytné rovněž pro férovou podporu distribuovaných úloh
    - Souběžné spuštění na více strojích
  - Známé přístupy vedou k velkému plýtvání zdroji
    - Řešeno blokací strojů, případně násilným ukončením úloh, které by ohrozily rezervaci
- Interaktivní úlohy
  - Potřeba okamžitého neplánovaného uvolnění zdrojů

Řešení pomocí virtualizace

- Stejný přístup pro advanced rezervace i pro interaktivní úlohy
  - V využití *preempce*
- Pozastavíme právě vykonávanou úlohu a zdroje poskytneme rezervované nebo interaktivní úloze
  - Virtualizace poskytuje nezbytné zapouzdření a ochranu
- Je možné rovněž využít migraci
  - Úloha potřebuje uzel s konkrétní vlastností, na kterém běží jiná úloha, kterou je možno přesunout na jiný generický uzel
- Další využití
  - Scavenger (nízkoprioritní) úlohy

Virtualizované plánování

- Systém Magrathea
  - Rozšíření standardních plánovačů o koncept virtuálních strojů
    - Nové stavy
    - Přepínání mezi virtuálními počítači na jednom uzlu

Denemark, Ruda, Matyska: Magrathea—Grid Management Using Virtual Machine, CGW06, 2007

Ruda, Denemark, Matyska: Scheduling Virtual Grids: the Magrathea Systém. Int. Workshop on Virtualization Technology in Distributed Computing, ACM DL, 2007

Bezpečnost versus jednoduchost použití

- Bezpečnostní požadavky někdy v rozporu s potřebami uživatelů
  - Bezpečnostní záplaty mohou měnit chování systému (numerická stabilita)
  - Starší aplikace nemusí být kompatibilní s moderními autentizačními metodami
  - V yladění/konfigurace systému pro konkrétní aplikaci vyžaduje administrátorské oprávnění
  - Komplikace při stavění MPI a podobných paralelních úloh
    - Timetout při autentizaci může zhroutit celý výpočet

Virtualizované řešení

- Úlohu zapouzdříme do virtuálního počítače
  - Tomu omezíme přístup do sítě
  - Uživatel ani s administrátorským oprávněním nemůže tato omezení překonat
- Virtuální počítač může spouštět i zastaralé (nezáplatované) verze operačních systémů a aplikací
  - V případě interaktivního přístupu vytvoříme chráněnou bránu,která propustí provoz mezi počítačem uživatele a virtuálním počítačem
- Je možné definovat celé virtuální clustery a uživateli zajistit chráněný přístup
  - Clustery mohou běžet vlastní/vyladěné verze operačního systému uživatele

Virtualizace sítě

- Podpora vysokorychlostní komunikace mezi virtuálními počítači na rozlehlé sítí
  - Nezbytné např. pro MPI úlohy
- Virtualizace na tzv. Level 2 (v podstatě virtuální Ethernet)
  - Umožňuje spouštět současně virtuální clustery se stejnými IP adresami
    - Vhodné zejména pro starší aplikace, které nemají podporu dynamické změny IP adres
  - Podporuje stejně IPv4 i IPv6
- Řízené plánovačem
  - Po počátečním nastavení není třeba zásah síťových administrátorů
- Aktuálně na páteři CESNET2 mezi Brnem, Prahou a Plzní

Antos et al: VirtCloud: Virtualising Network for Grid Environments, AINA 2009, submitted

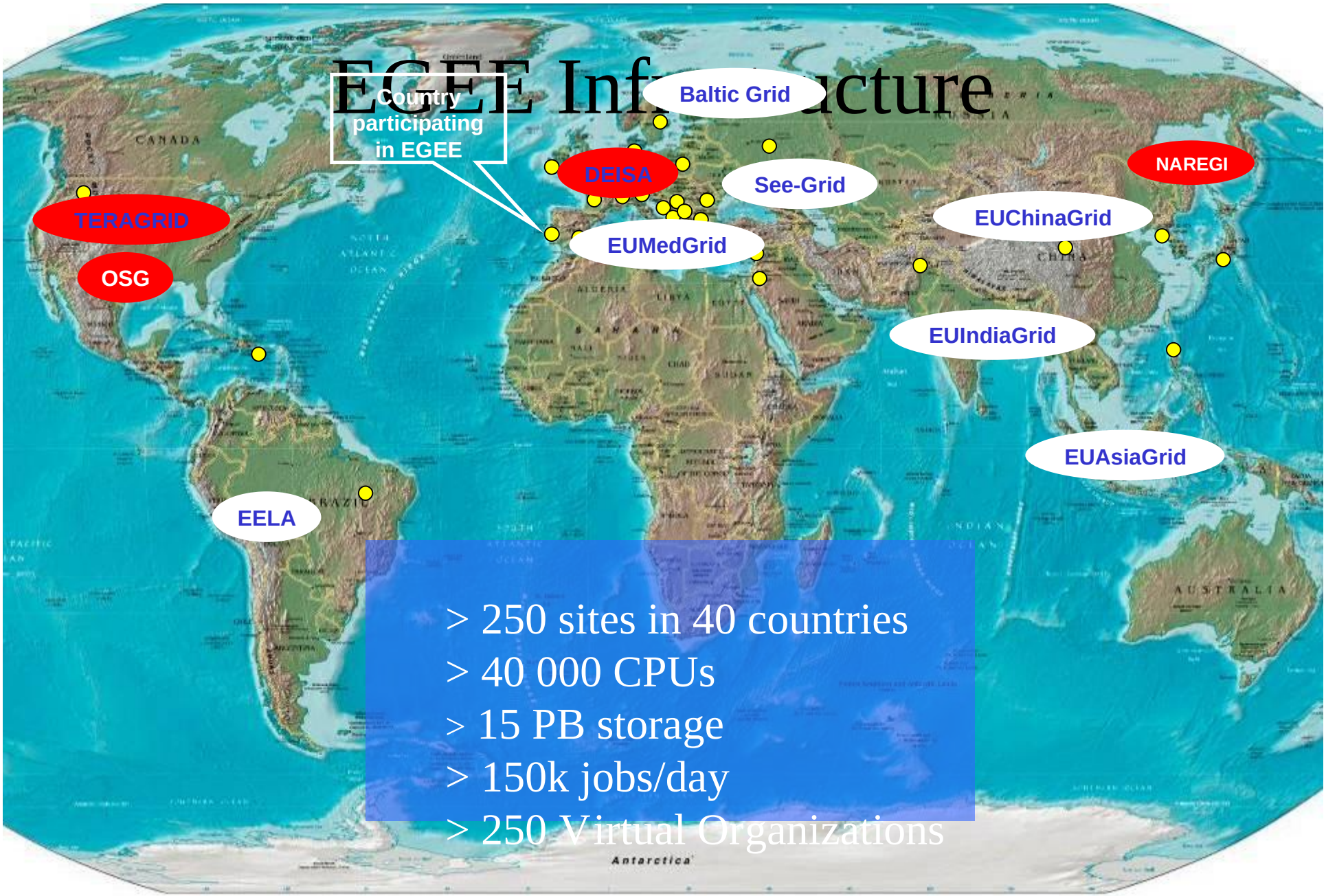
Distribuované výpočetní infrastruktury – hlavní reprezentanti

- Evropa
  - EGEE – Enabling Grids for E-sciencE
- USA
  - TeraGrid
  - OSG – Open Science Grid
- Asie
  - Japonsko: Naregi
  - Čína: ChinaGrid (IPv6)

EGEE

- Série dvouletých EU projektů
  - Aktuálně EGEE III
  - Průměrný rozpočet cca 50 MEuro, příspěvek Unie kolem 35 MEuro
- Zajištění provozu
- Podpora uživatelů a aplikací
  - Pomoc při „gridifikaci“
- Součástí vlastní vývoj middleware – gLite
  - V podstatě kompletní řešení výpočetní a úložné infrastruktury
  - LB a JP služby výsledkem našeho zapojení
    - Instalovány na desítkách míst v produkční infrastruktuře

EGEE Infrastructure



EGEE – organizace

- Hlavní koordinátor: CERN
- V EGEE II více jak 90 partnerů, EGEE II stále cca 50 partnerů
  - V EGEE III z každé země jen jeden akademický partner
    - Realizováno mechanismem JRU (Joint Research Units)
    - Za ČR sdružení CESNET
  - Regionálně sdruženo do *federací*
    - ČR součástí Středoevropské federace (Rakousko, Polsko, Slovensko, Maďarsko, Slovinsko, Chorvatsko)
    - Zastupuji federaci v rámci Project Management Board

EGI DS

- European Grid Initiative Design Study
- Projekt zaměření na nalezení vhodného organizačního i finančního modelu dlouhodobě udržitelné celoevropské gridové infrastruktury
- V podstatě na zakázku Evropské komise
  - Definovat prostředí a funkce na národní i EU úrovni, navrhnout postupy a připravit vlastní realizaci (včetně doporučení grantovým agenturám i EU ohledně financování)
  - Pouze 9 partnerů, ale ustaven Policy Board, v němž jsou zástupci všech členských států EU (i dalších zemí)
- Postaven na konceptu Národních gridových iniciativ/infrastruktur
  - Analogie NREN (National Research and Educational Network operator)
- Zahájen 1. srpna 2007
  - Od 1. července 2008 jsem koordinátorem projektu

Česká republika – projekt MetaCentrum

- Součást výzkumného záměru sdružení CESNET
- Tři hlavní uzly
  - Plzeň, Praha, Brno
- Přes 1300 jader a téměř 100 TB úložné kapacity
  - Plus 2*200 TB v páskových knihovnách
- Široké spektrum aplikací
  - Převažuje výpočetní chemie a obecně univerzitně orientované výpočetní vědy
  - Přes 150 uživatelů

Shrnutí

- Distribuované systémy představují budoucnost prostředí, které dnes nazýváme Internet
  - Sdílení a vzdálený přístup ke zdrojům, sdílení a zpřístupnění dat, informací i znalostí, stále se rozšiřující nabídka nových služeb, podpora spolupráce, …
  - E-Infrastruktura je ve stále větší míře akceptována jako nezbytná základ výzkumného prostoru bez ohledu na konkrétní vědu a její orientaci
- Konstrukce reálně použitelných distribuovaných systémů je stále velmi složitý problém
  - Peer to peer sítě sice spojují miliony uzlů, ale jejich efektivita je příliš nízká, případně nejsou jasné bezpečnostní důsledky (např. skype)
- Rozsáhlé systémy příliš složité pro jednoduché analytické zpracování
  - Neexistence globálního stavu
  - Analogie se systémy v živé přírodě
  - Problém reprodukovatelnosti experimentálních výsledků

Vlastní práce

- Kombinace návrhu, budování a provozu rozsáhlých distribuovaných systémů
  - Nejrůznější aspekty tohoto prostředí
  - Sítě, plánování, middleware, bezpečnost, uživatelé, aplikace
  - Distribuovaný systém jako základ prostředí pro spolupráci
- Fokus různorodých výzkumných aktivit
  - Od základních konceptů architektur až po reálnou produkční implementaci
  - Rychlý přenos výsledků do praxe na národní i mezinárodní úrovni
- Výzkum v této oblasti znamená kombinaci teoretických i experimentálních přístupů a simulací
  - Jedná se již dlouhodobě o jednu z největších výzev Informatiky

Děkuji za pozornost

Dotazy?