Big Data Applications on Multi-Clouds
An Introduction to the MELODIC Project

Thomas Dreibholz (托马斯博士)  29 December 2017
Simula Research Laboratory, Norway  Haikou, China
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731664.

Motivation
An Introduction to Cloud Computing
The MELODIC Project
Use Cases
Conclusion
From PC to Cloud Computing

• In former times:
  – Powerful desktop PC (personal computer)
  – Fast CPU, large harddisk(s)
  – Disadvantage: expensive and maintenance-intensive

• Today (or in near future):
  – Laptop or tablet PC
  – Energy-efficient components (battery-powered)
  – Cloud applications, e.g.:
    • Software as a Service, web mail, calender, office, photo management, ...
    • Storage space (DropBox, ...)

A „normal“ user mainly works in the web browser!
Smartphone and Cloud Computing

• Smartphone
  - Small storage space (or expensive)
  - Not extensible (e.g. SD card slot)
• Cloud connection
  - Storage space
    • Pictures, videos, music
    • Documents
  - Applications
    • Computation-intensive applications in the cloud (z.B. voice recognition)

Cloud connection is a major component of smartphones!
The Challenges

Cloud Computing is convenient with good network connectivity!

- Challenge: network communication
  - Bandwidth
  - Latency
  - Mobility

How does Cloud Computing work? (and, by the way, what is the „Cloud“?)
What is the „Cloud“?

There is no cloud
it’s just someone else’s computer

Image sources:
CompuServe advertisement 1988
Chris Watterston
Cloud Computing – A Technical Overview

• From bottom to top
  − Hardware
  − **Virtualisation**
  − Management Frameworks
  − Applications
  − Software as a Service (SaaS)
Hardware in External Data Centres

• User's Local PC:
  – Low utilisation
  – Main task: do nothing!

• Idea: many computers, for very many users!
  – Computers in a data centre
  – Usage by many users
    • Temporal distribution of the users
    • High utilisation
    • Low costs

Multiple customers per computer?
Virtual Machines: Basics

- Virtualisation: "sharing" of hardware by using a hypervisor
- Virtual Machines (VM):
  - Own virtual hardware (CPUs/cores, memory, network, storage)
  - Own operating system (OS) within the VM
  - Hypervisor may emulate "standard" hardware → no problems with drivers
Virtual Machines: Different Variants

- **Hypervisor**
  - **Type 1**: Runs directly on the hardware
    - e.g. VMware ESXi
  - **Type 2**: Part of “normal” operating system
    - KVM (Kernel-based Virtual Machine) in Linux,
    - VirtualBox, etc.

- **Para-virtualisation**
  - Instead of emulating virtual hardware: just pass through OS calls
  - Requires OS adaptation/drivers

Full operating system in every VM → overhead
Containers

- Container
  - One OS kernel, multiple restricted areas (containers)
  - Each container "sees" only
    - Own processes
    - Own virtual network interfaces
    - Own view on file systems
    - ...
  - OS manages containers – has access to everything

- Widespread implementations:
  - Linux Containers (LXC) and FreeBSD Jails

Low overhead, but restriction to the same kernel
Management Frameworks

VMs and containers are convenient! But how to manage them?

• Orchestration
  – Instance management:
    • create, remove, migrate, backup, restore, configure, ...
  – Service monitoring:
    • Is everything running, or are there problems somewhere?
• Well-known management frameworks:
  – OpenStack
  – Docker
  – Kubernetes
  – and many more!
Applications for VMs and Containers

• „I need a web-server with database“
  – New container from template
    • Ubuntu Server + Apache + PostgreSQL
    • …
  – VM with Windows + IIS + Oracle

• „I need a FreeBSD system for kernel tests“
  – FreeBSD-VM from template
  – Snapshot/restore possible
  • …

The „normal“ user just wants to use his application!
Software as a Service (SaaS)

- Instead of own VMs/containers:
  - Subscribe to a complete service (e.g. as web application)
  - Provider takes care for everything

- Examples:
  - Microsoft Office 365
  - Google Apps (Mail/Drive/Photos/...)
  - Yahoo (Mail/Flickr/...)
  - GitHub, BitBucket, ...

Very simple and convenient for the “normal” user!
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731664.

Social innovations in modern era increasingly rely on our capacity to process large datasets.

- **Digital Universe** – Expected to grow to 44 ZiB in 2020.

- **Internet-of-Things** – 30.7 billion devices by 2020.

- **Social Networks and Multimedia** – 500 million tweets per day; 510,000 comments and 136,000 photos per second on Facebook!

- **Biological Data** – Vast amount of data available for researchers, 1000 Genome project, 100,000 Genome project, TiB to EiB!

---

**Big Data**

- **Volume**
- **Velocity**
- **Veracity**
- **Variety**
- **Value**

Big data needs big resources!

- $1 \text{ TiB} = 1024^4 = 1,099,511,627,776$ bytes
- $1 \text{ EiB} = 1024^6 = 1,152,921,504,606,846,976$ bytes
Currently, users are restricted to static deployment choices for data-intensive applications.

- Cost Effectiveness
- Management Flexibility
- Resource Utilization
- Privacy and Confidentiality

- Cost Effectiveness
- Management Flexibility
- Resource Utilization
- Privacy and Confidentiality

- Cost Effectiveness
- Management Flexibility
- Resource Utilization
- Privacy and Confidentiality

- Cost Effectiveness
- Management Flexibility
- Resource Utilization
- Privacy and Confidentiality

- Vendor Lock-In
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731664.

Melodic is infrastructure-agnostic, support full life-cycle of data-intensive applications.

- **APPLICATIONS**
  - Unified Cloud View to Serve the User Applications
  - Big Data Analytics Enterprise real-time processing

- **TOOLBOX**
  - DATA PROCESSING
    - Spark
    - OpenStack

- **DATA SOURCES**
  - On Site
  - On Private Cloud

**Public Cloud**
- **MELODIC**
  - Access Management
  - Security Policies
  - Non-functional Requirements
  - Resource Management Interfaces
  - Monitoring Interfaces

**Private Clouds(s)**
- **White-Box approach**
- **Grey-Box approach**

**On Site**
- **On Private Cloud**
- **On Public Cloud**

- **Location-aware data-intensive processing on federated clouds**
- **Resource optimization in hybrid multi-cloud environments**
- **Trust, security and privacy in geo-distributed clouds**

**EU Projects targeting isolated MELODIC objectives**
- CYCLONE
- BEACON
- ALIGNED
- DICE
- CACTOS
- SSICLOPS
- MUSA
- PaasWord
- ESCUDO
- CLARUS
A Complete solution for data-intensive applications

- **Transparent** deployment and execution of data-intensive applications on multi-clouds
- **Holistic data** life-cycle management
- **Runtime adaptation** and automatic elasticity for cloud applications
- **Secure and privacy-aware** data access
- **Optimal usage** of distributed private infrastructures with federated clouds
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731664

Big Data Cloud Made Easy!

Melodic calculates **best** multi-cloud options for your applications

Automatic deployments and Adaptation

Improved **performance** at lower **costs**!
BIG IDEA: AVOID VENDOR LOCK-IN

This is what we believe in:

- **Cloud-future** with competition not monopoly.
- **Choice**, change & opportunities.
- **Multi-cloud** complexity made simple and manageable.
OPEN SOURCE
OPEN STANDARDS

We believe that open source projects promote innovation faster than proprietary solutions.
JUST TELL MELODIC WHAT YOU NEED

Specify your needs once and forget about differences between cloud operators, multiple admin panels and other headaches.

- provisioning? deployment?
- scalability? service level?
- jurisdiction?
- cost concerns?
- monitoring?
Melodic will be demonstrated with four selected use-cases (covering different deployment aspects and user requirements)

- **Genome Analysis**
  - Strict Data Confidentiality, Performance Optimization, Cost Effectiveness

- **Road Traffic and People Flow Monitoring**
  - Real-time Processing, On-Demand Processing, Geo-dispersed Big Data, Data privacy

- **Secure Document Management**
  - Data Life-cycle Management, Performance Optimization, Commercial Exploitation

- **Marketplace for Data-Intensive Apps**
  - Transparent Deployment, Runtime Adaptation, Automated Elasticity
Use Case: Genome Analysis

Białystok University provides application prototype enabling robust approach for the discovery of synergistic variables in biological datasets, with a main focus on data from gene expression studies and genome-wide association study (GWAS).

Melodic enables:

- Utilize cloud computing processing power and scalability
- Minimize data processing costs
- Use innovative solutions (like GPUs) to speed up data processing capacity
MELODIC CASE
STUDY: CE-TRAFFIC

Melodic helps CE-Traffic calculate road traffic and people flow information

- Effortless switching between cloud providers minimizes cost.
- Big-data-cloud complexity made easy.
- On-time results.
Big Data Cloud Made Easy!

Melodic calculates **best multi-cloud options** for your applications

Automatic deployments and Adaptation

Improved **performance at lower costs!**
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731664.

Thomas Dreibholz (托马斯博士)
dreibh@simula.no

Date:
29 December 2017

www.melodic.cloud
facebook.com/MelodicCloud
twitter.com/melodic_cloud
linkedin.com/MelodicCloud
slideshare.net/MelodicCloud
Merry Christmas!

MELODIC
Multi-cloud Execution-ware for Large-scale Optimised Data-Intensive Computing

MELODIC / NorNet Core
Simula Research Laboratory
http://www.melodic.cloud | https://www.nntb.no