Rethinking Computer Architecture: The BSC-Microsoft Research Centre

Osman S. Unsal
Senior Investigator, BSC-MSRC

DAC (UPC): High Performance Computing

Computer architecture:
- Superscalar and VLIW
- Hardware multithreading
- Design space exploration for multicore chips and Hw accelerators
- Transactional memory (Hw, Hw-assisted)
- SIMD and vector extensions/units

Benchmarking, analysis and prediction tools:
- Tracing scalability
- Pattern and structure identification
- Visualization and analysis
- Processor, memory, network, system

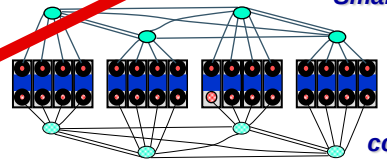
Programming models:
- Scalability of MPI and UPC
- OpenMP for multicore, SMP and ccNUMA
- DSM for clusters
- CellSs, streaming
- Transactional Memory
- Embedded architectures

Grid/Cloud

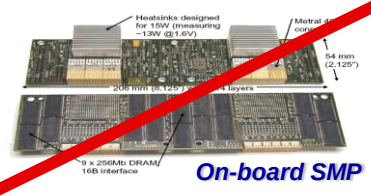
Future Petaflop systems

Large cluster systems

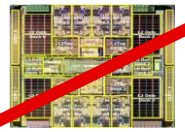
Small DMM



cc-NUMA



On-board SMP



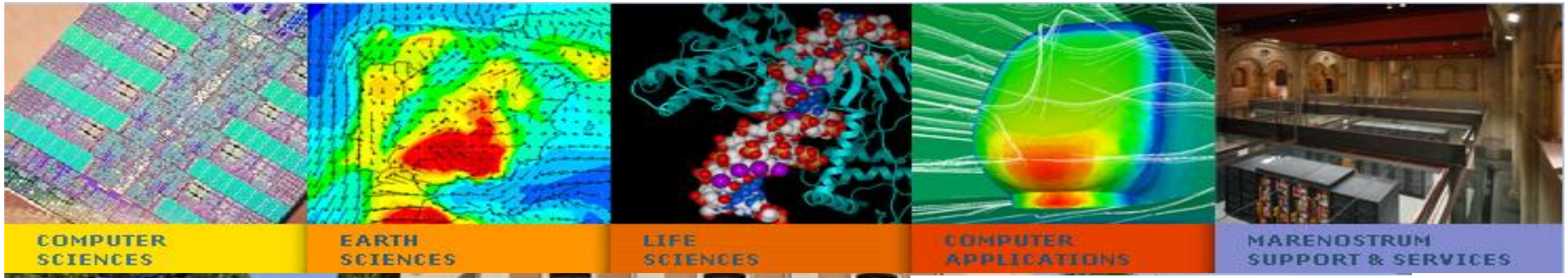
Chip

Grid and cluster computing:
- Programming models
- Resource management
- I/O for Grid
- Cloud Computing

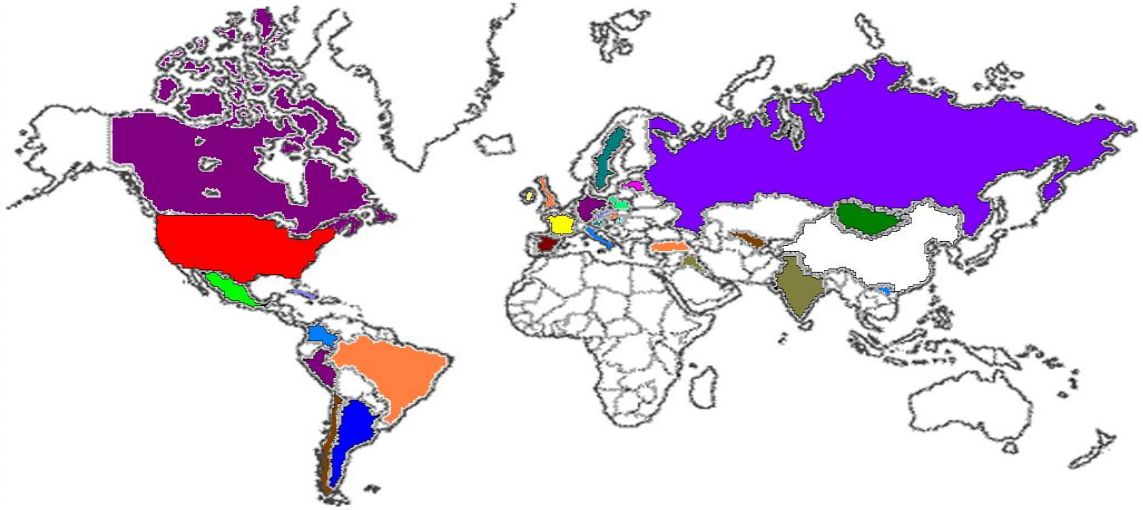
Algorithms and applications:
- Numerical
- Signal processing

Operating environments:
- Autonomic application servers
- Resource management for heterogenous workloads
- Coordinated scheduling and resource management
- Parallel file system scalability

The Barcelona Supercomputing Center



300 people from 24 different countries
(Argentina, Belgium, Brazil, Bulgaria, Canada, Colombia, China, Cuba, France, Germany, India, Iran, Ireland, Italy, Jordania, Lebanon, Mexico, Pakistan, Poland, Russia, Serbia, Spain, Turkey, UK, USA)



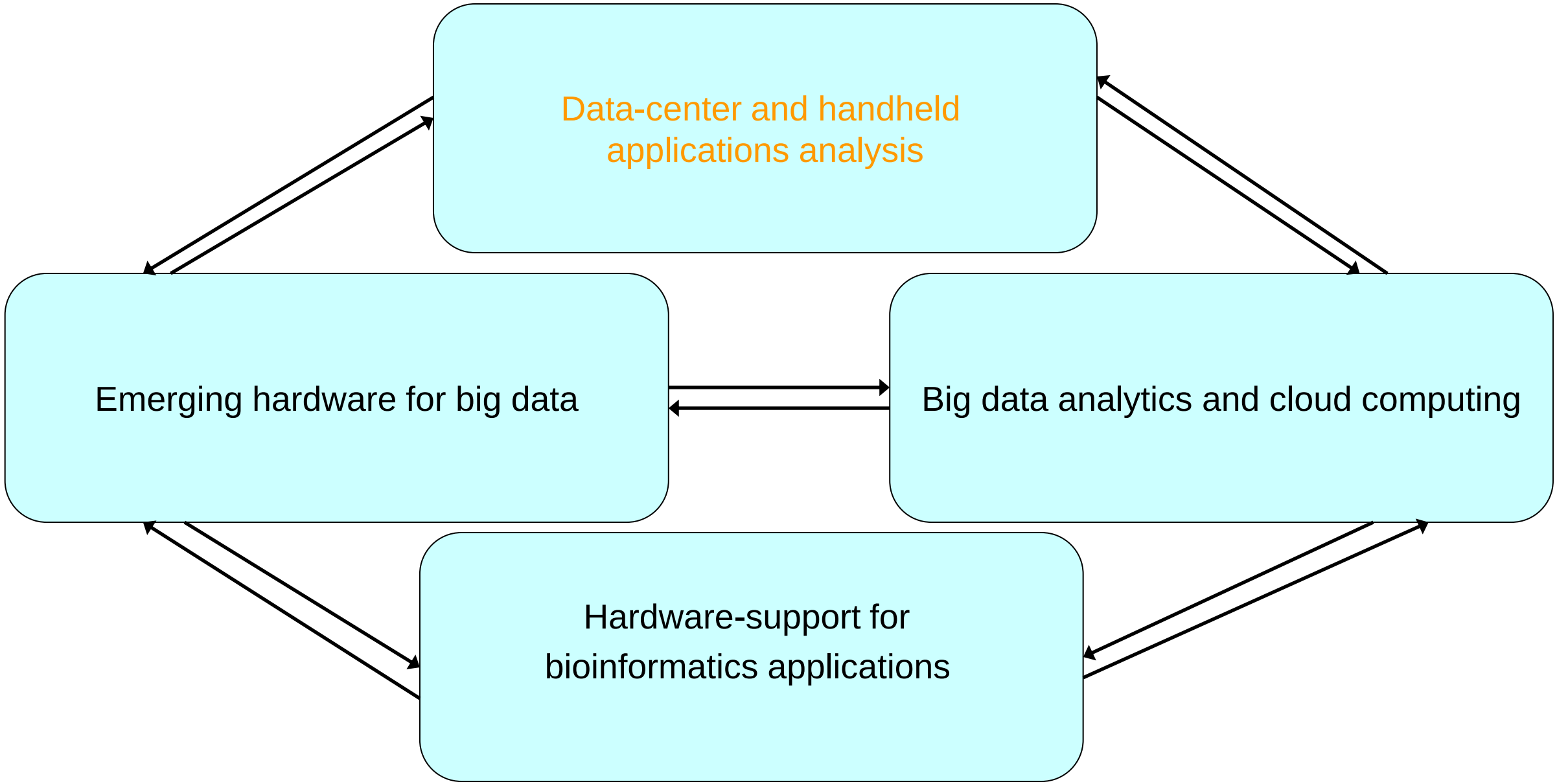
BSCMSR Centre: general overview

- Microsoft-BSC joint project
  - Combining research expertise
    - Computer Architectures for Parallel Paradigms at BSC: computer architecture
    - Microsoft Research Cambridge: programming systems
  - Kickoff meeting in April 2006
  - BSC Total Effort:
    - Very young team! 3 Senior BSC researchers, 18 PhD + 1 engineer
  - Support and part-time involvement of senior researchers and faculty from UPC
- BSCMSRC inaugurated January 2008

Outline

- Current Research Directions
- Past Research
- BSC Collaboration with Latin America

Current Research Directions



The big consolidation

- Processors for the mobile, server and desktop market segments are converging:
  - All segments require RMS applications
  - Everything will run on a cloud
  - Witness Intel: unique core for all segments
- Traditional form factors are shrinking:
  - The palmtop
- Everything must run everywhere

Everything must run everywhere

Gyroscope

Handwriting Recognition

Machine Learning

Cameras

Image Recognition

Data Mining

Digital Health

Low-power and High Performance

Immersive games

Speech Recognition

Automatic composition

Augmented Reality

Virtual Reality

Smell, Proximity, Haptic sensors

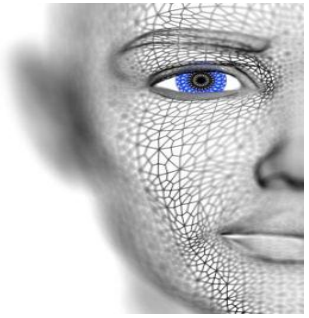
3D HUD

Speech Synthesis

Automatic GPS tagging

Profiled applications

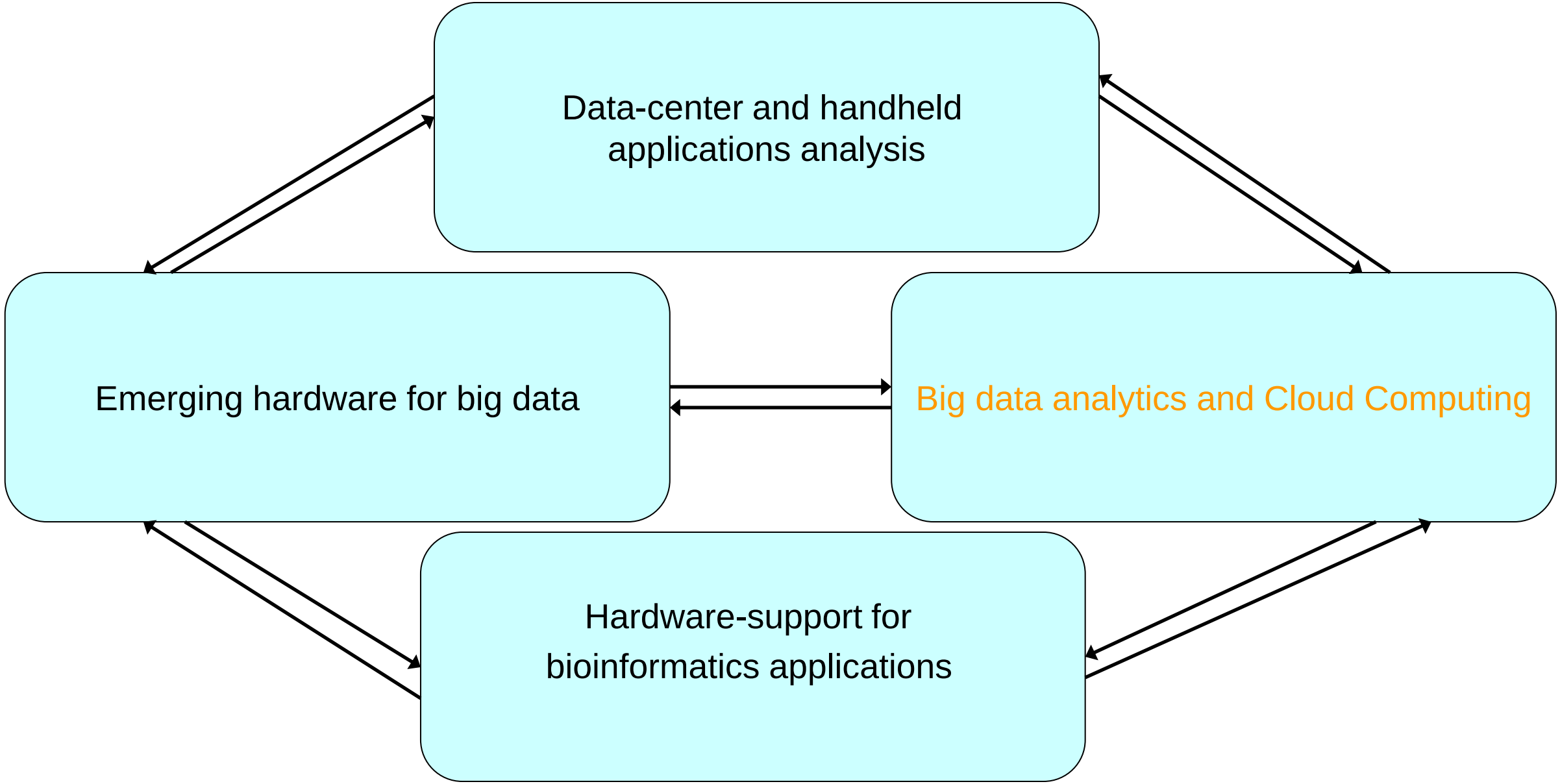
- Sphinx3 – Speech recognition
- Festival – Speech synthesis
- SPEC2000/CSU – Face Recognition
- San Diego – Computer Vision
- C-Store – Column-Store DBMS
- NU-MineBench – Data Mining



Profiling results

- Large data level parallelism exist
  - Significant % of execution time
- Even if application not vectorizable
  - Can be transformed to leverage vectors with reasonable effort targeting:
    - small code sections
    - impact execution time significantly

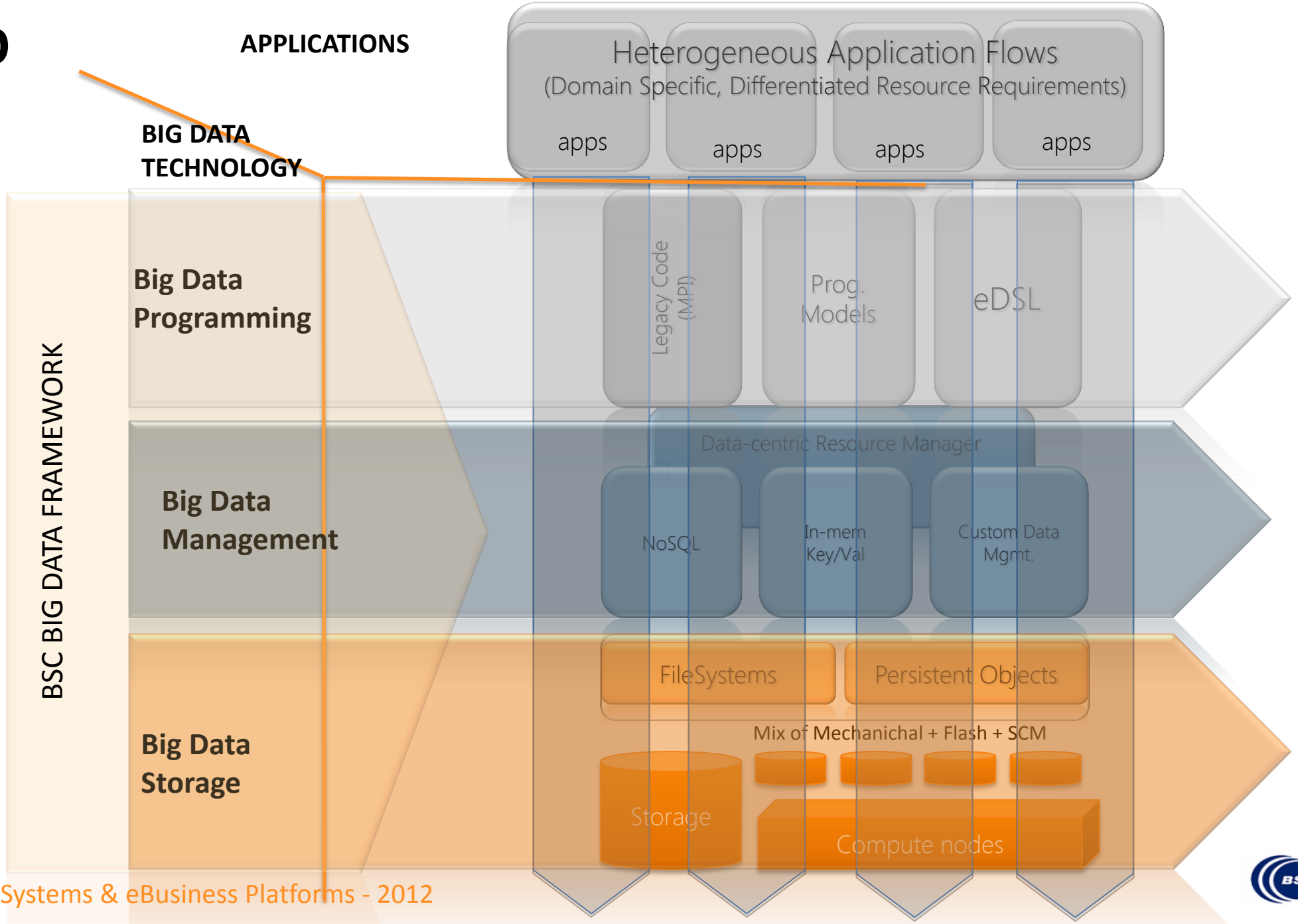
Current Research Directions



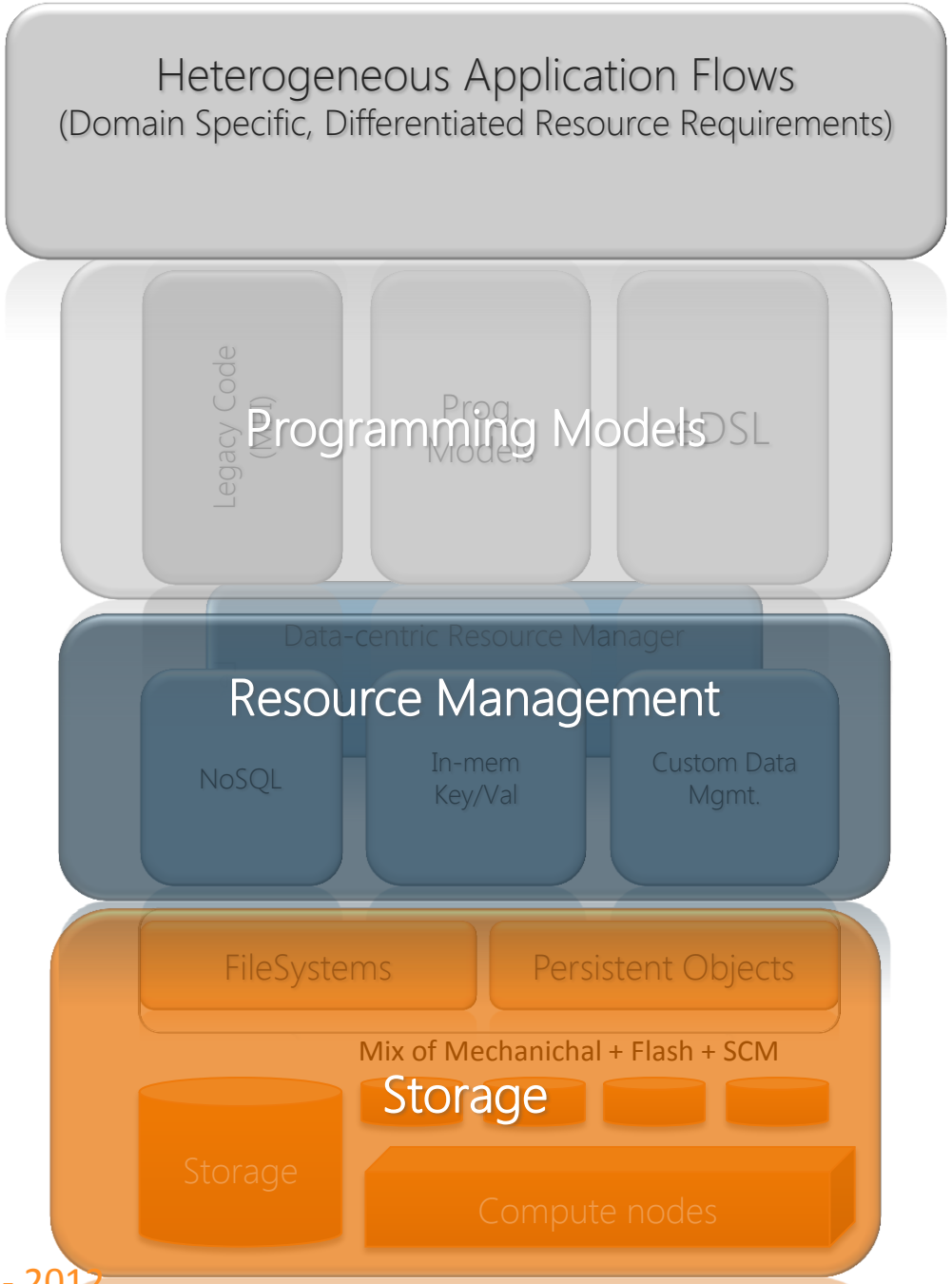
Big Data Analytics and Cloud Computing

- Severo-Ochoa project
- AXLE Project
- Venus-C Project

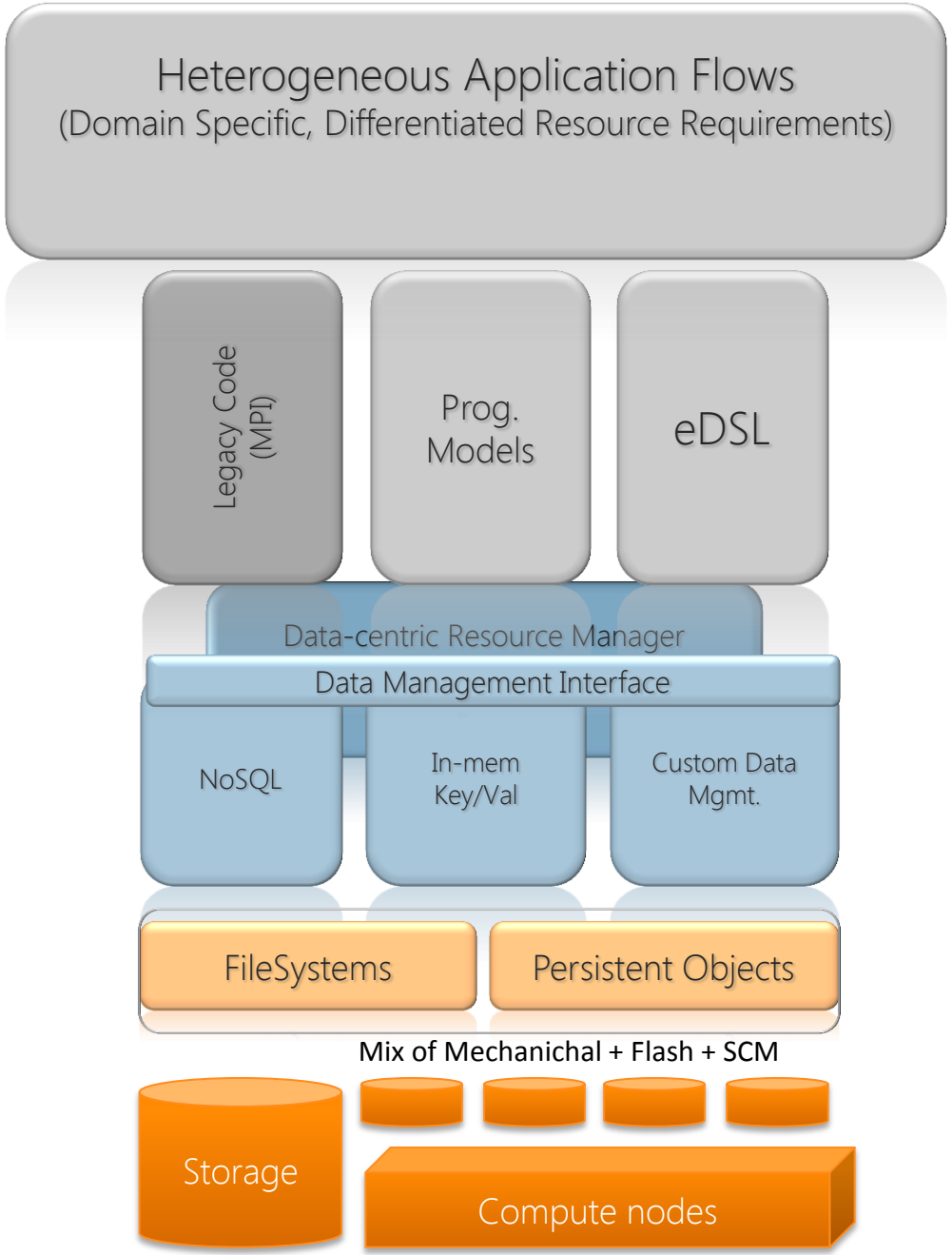
CS-SO Stack



Simplified Stack vision



CS Stack



Big Data Analytics and Cloud Computing

- Severo-Ochoa project
- AXLE Project
- Venus-C Project

Big Data Analytics:Target problem

- The future brings us extremely large volumes of data
- We need better performance
- We need better security, privacy, and auditing of data accesses

PostgreSQL DB

- Enterprise-grade
- Widely used
- Today handles 100s of GB of data (up to 1 TB)
- Scales horizontally and vertically
- Has many extensions: custom data types, analytics, etc.

Security, Privacy, and Audit

- Mandatory Access Controlled Row-Level Security
  - Currently on the level of tables & views
- Automatically create and update the database Security Policy
  - Currently, security policies are mostly written manually (error-prone)

Scalability Engineering

- Dynamic compilation and optimization of searches (SQL & NoSQL)
- SQL and NoSQL on GPUs and FPGAs
  - Better performance and power efficiency
  - Parallel execution of a single query
  - Outboard sort and bitmap data
  - Automatic Partitioning
  - Proving the certain parts of tables need not be included as part of the query

Advanced Architectures

- Non-Volatile RAM: ferroelectric, magnetoresistive, phase-change memory
  - Can significantly simplify the implementation of DBMS
  - Reduce the latency
  - A rough estimation: 3-4 orders of magnitude faster execution
- 3D stacking
  - Reduce the latency
  - Reduce the power consumption
- Vector operations
  - Energy efficiency

Visual Analytics

- Visualization is important for understanding
- Extend the “Orange” data-mining system
  - We can’t simply put millions of points on the screen?
    - Overview first
    - Zoom and filter
    - Details on demand
  - Performance: execute most queries in DBMS
    - Better performance

Evaluation on Real Data

- Real medical data:
  - Important: covers 200,000 individuals in Netherlands
  - Highly secure: it's **real** data, of **real** people
  - Complex
  - Standardized
  - Large: 300 GB, we will extrapolate to 9 TB
- Will use BSC data (20TB) from Severo Ochoa database

Big Data Analytics and Cloud Computing

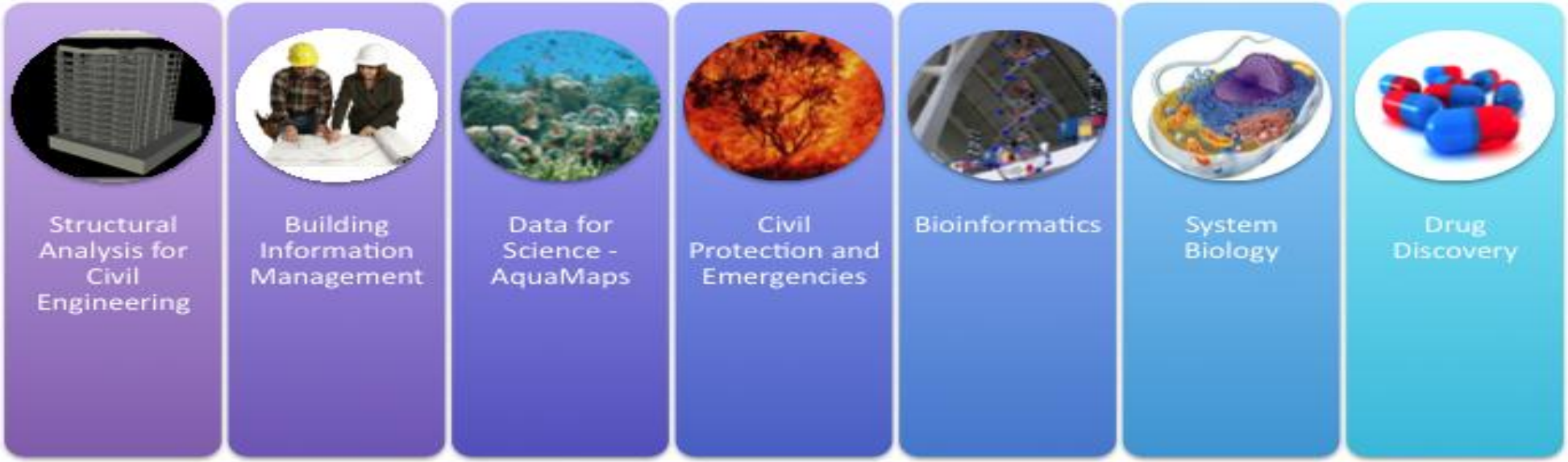
- Severo-Ochoa project
- AXLE Project
- Venus-C Project

Venus-C

- EU funded project
  - 15 partners, between them BSC and Microsoft Research
  - Microsoft contributes with Azure resources and manpower
- Goals
  - to create a platform that enables user applications to leverage cloud computing principles and benefits
  - to leverage on the state-of-the-art to bring on board early adopters quickly
  - to create a sustainable infrastructure that enables cloud computing paradigms for the research communities

Venus-C

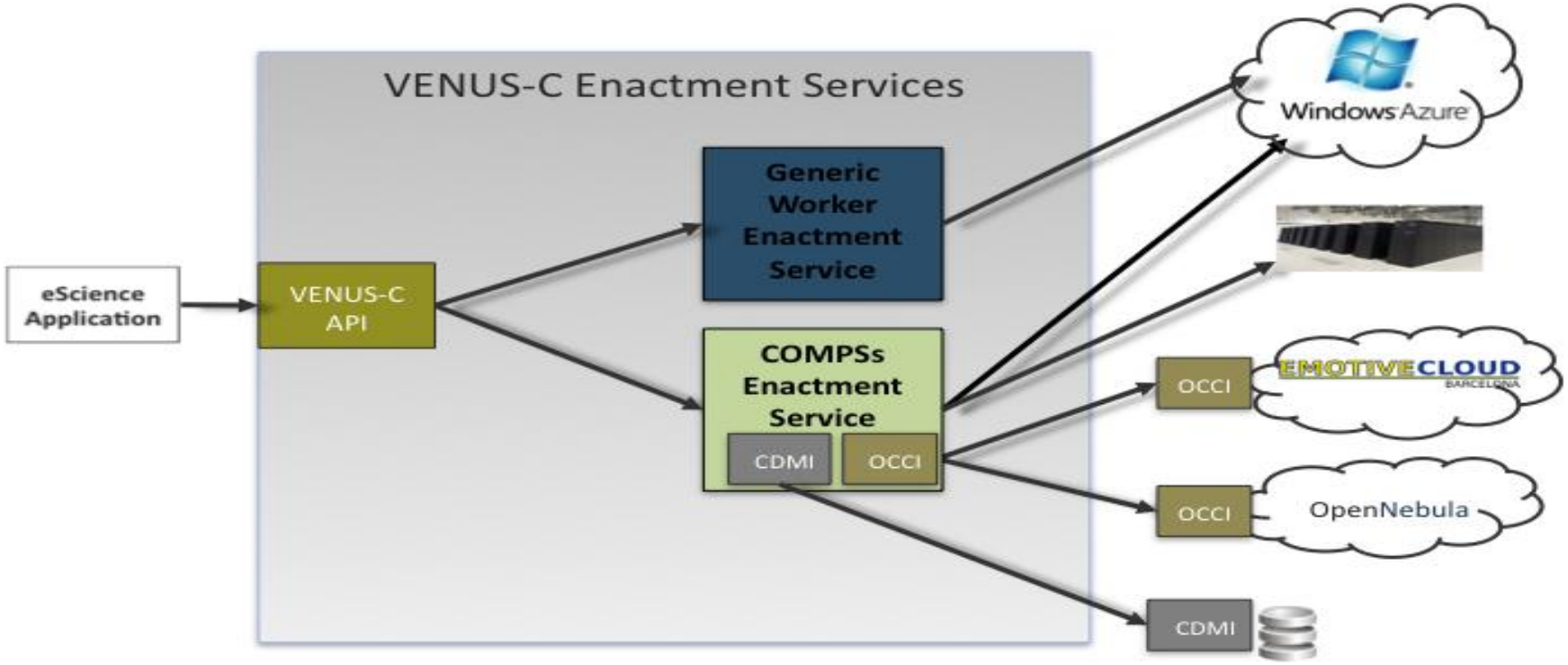
- Activities
  - Development of the Venus-C platform
  - Deployment of 7 scenarios applications in the platfom
  - Deployment of 15 pilots applications in the platfom



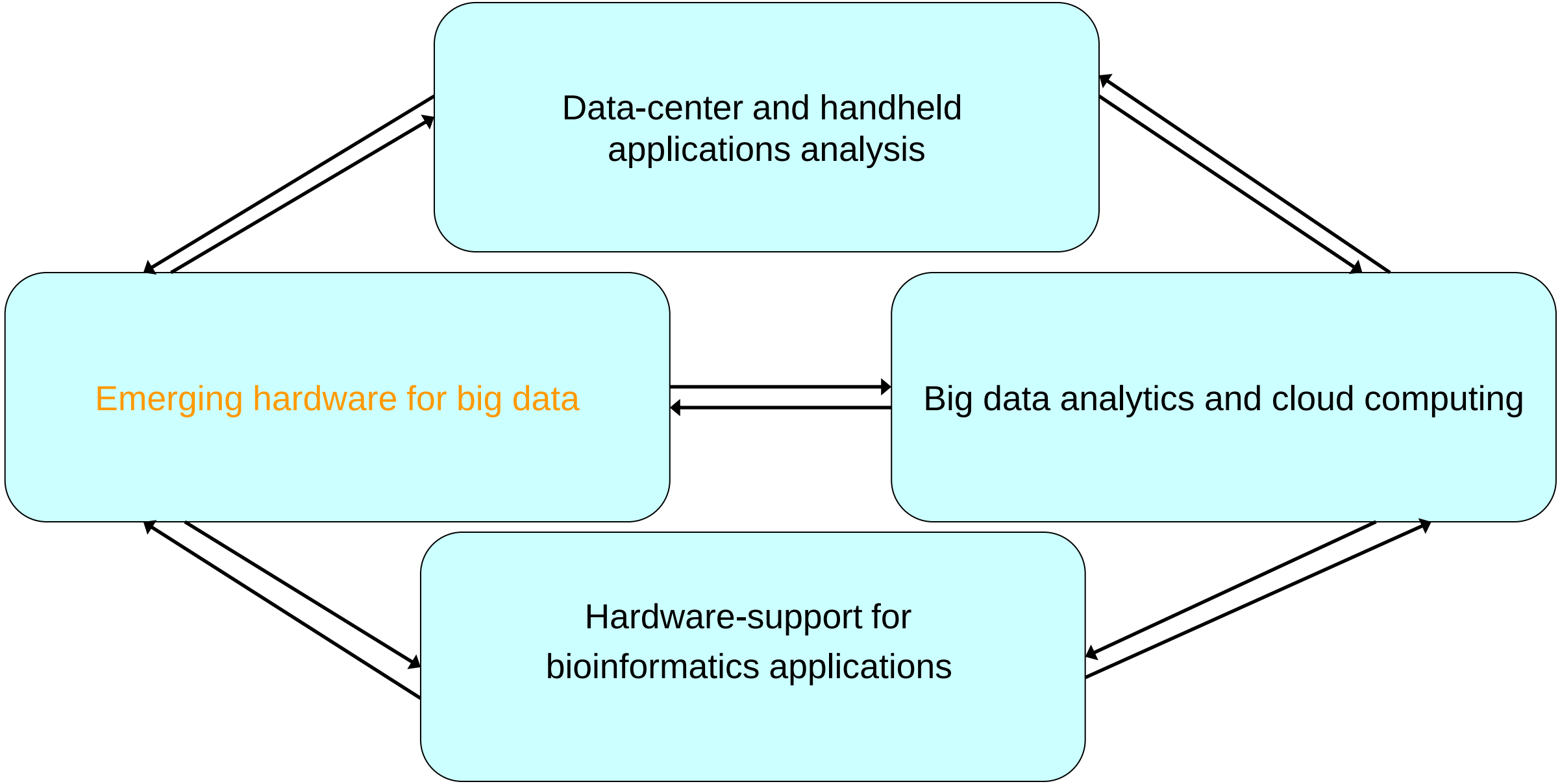
Venus-C

COMPSs (BSC) in VENUS-C allows the easy porting of scientific applications to Clouds

- Minimum modifications to original code
- Provides elasticity features
- Execution results demonstrate scalability
- Interoperability with different cloud providers, Azure between them



Current Research Directions



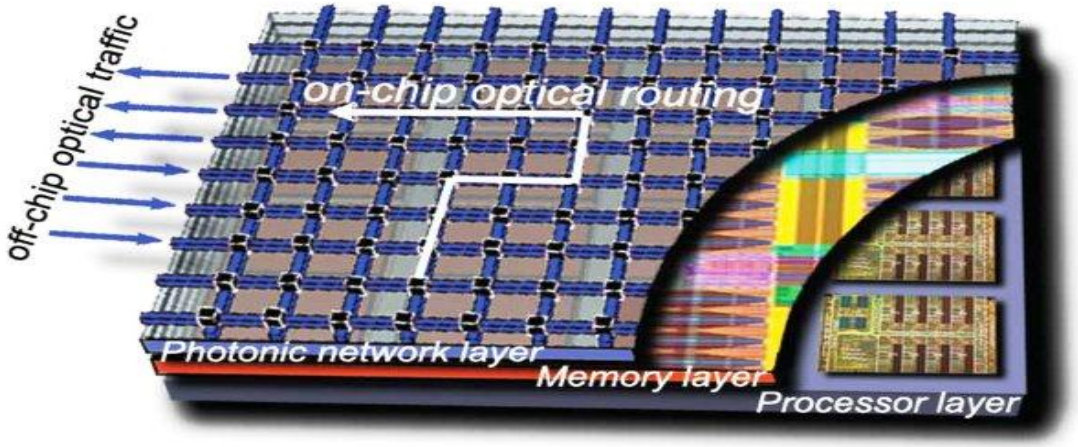
Emerging Hardware for Big Data

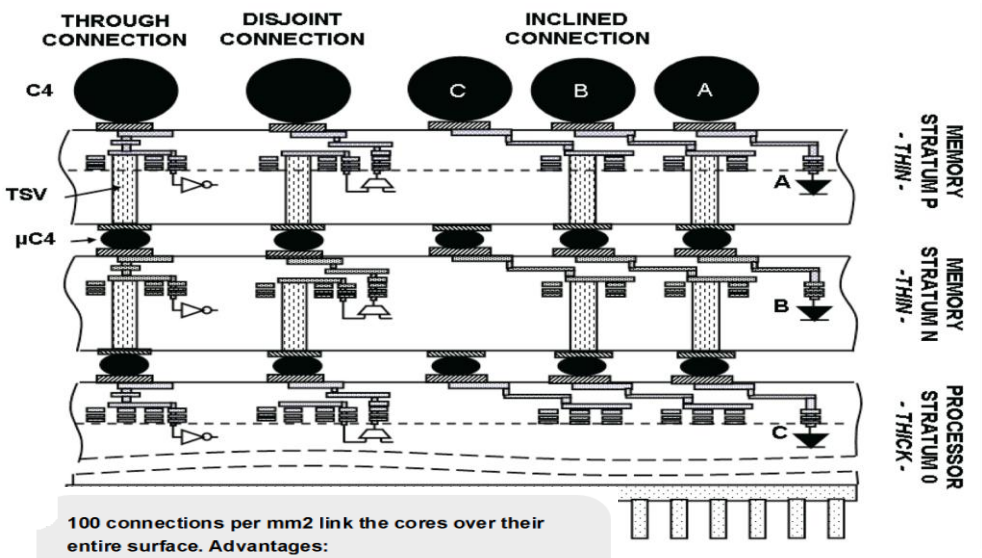
- Non-volatile+3D
- ParaDIME for Programming model support for radical energy savings
- Decision-support systems acceleration
- HW-support for sparse data

Emerging Technologies

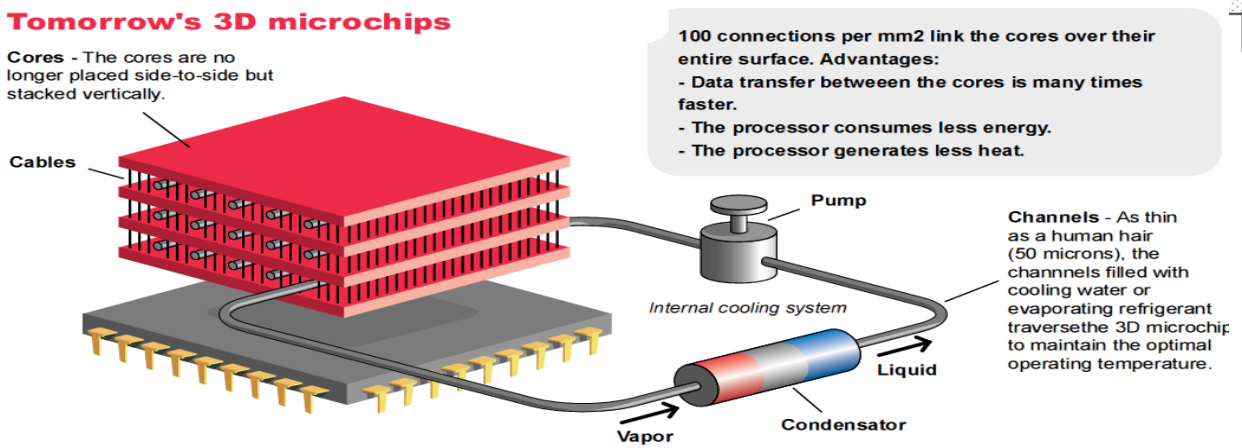
- Emerging technologies
  - 3D Stacking
  - Nonvolatile-RAM
- Develop an Architecture and prototype using these technologies

3D Stacking



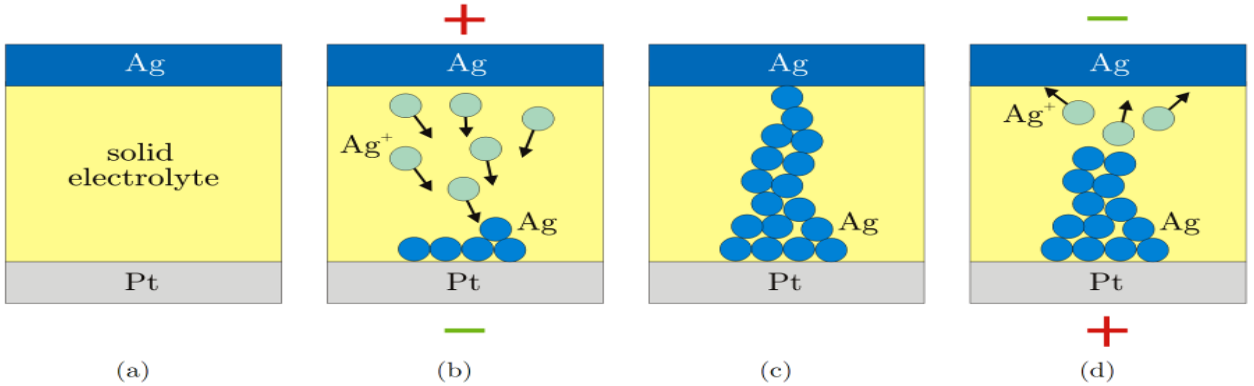


100 connections per mm2 link the cores over their entire surface. Advantages:
- Data transfer betweeen the cores is many times faster.
- The processor consumes less energy.
- The processor generates less heat.



Nonvolatile-RAM

- Non Volatile Memory
- Very low power devices
- Compatible with CMOS technologies



CB-RAM

	CMOS		NON-CMOS		
	SRAM	DRAM	PC-RAM	STT-RAM	CB-RAM
Data Retention	N	<s	>10y	Ms to 10y	>10y
Energy Write(J/ns/bit)	$<10^{-15}$	$<2*10^{-15}$	$<10^{-12}$	$<10^{-13}$	$<10^{-15}$
Refresh	None	Yes	None	Maybe	None
Leakage	Yes	Yes	None	None	None
Read Voltage (V)	<1	<1	3	<0.5	<0.2
Write Time (ns)	<1	50	<50	<5	<5
Cell size (F^2)	50-120	6-10	6-12	4-20	5-8
Feature Size (nm)	12	25	5-10	6	5-10
Endurance(write cycles)	10^{16}	10^{16}	10^{9}	$>10^{15}$	$>10^{15}$
Multilevel (MLC bits)	1	1	4	1	>4

Goals

- Integrate CB-RAM with 3D stacking to work as universal memory
  - Replacement of the memories caches, allowing to have multi-gigabytes bytes caches
  - Develop new coherence protocols for this memory hierarchy
- Using CB-RAM in the main core architecture, not only as replacement of caches and memory subsystem
  - Develop CAMs based on CB-RAM to be used in the core
  - Research in the tradeoff between write time, endurance, data retention and write energy

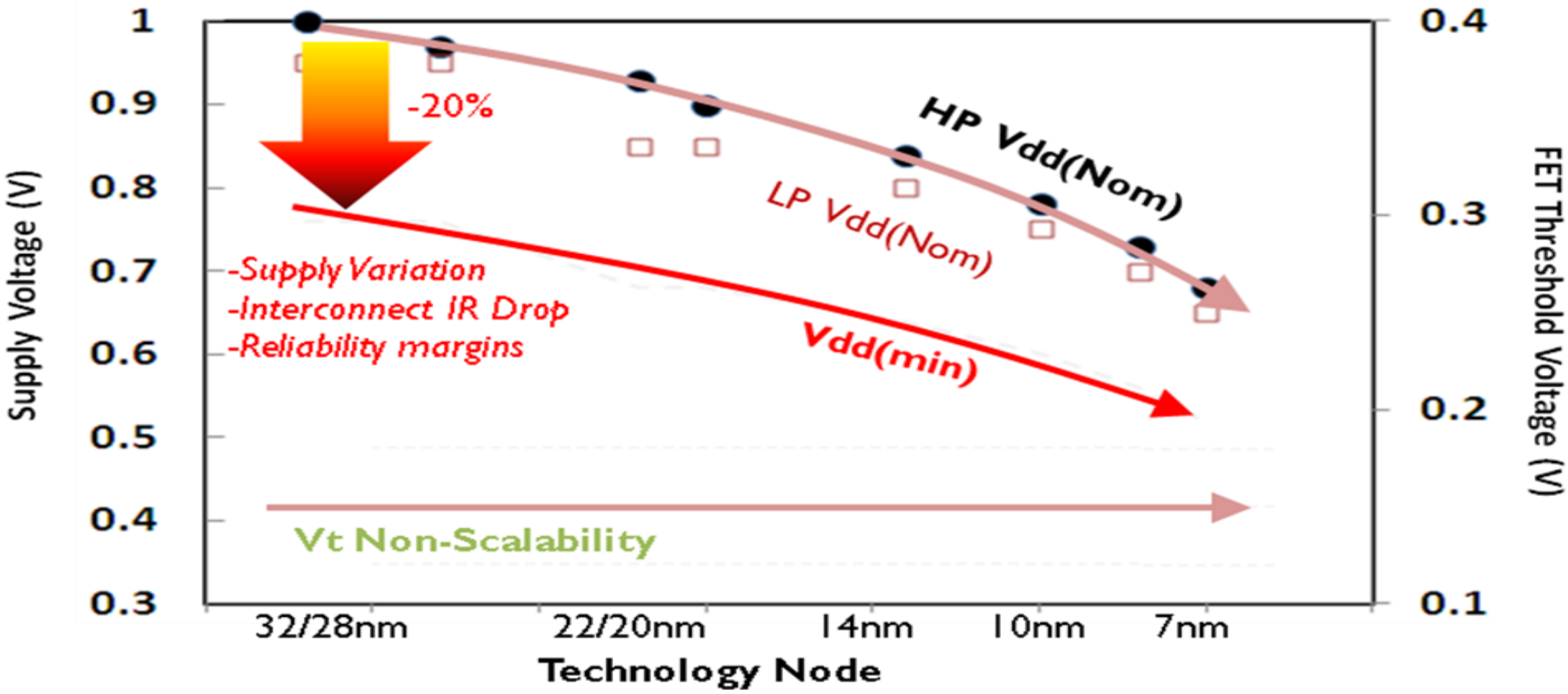
Emerging Hardware for Big Data

- Non-volatile+3D
- ParaDIME for Programming model support for radical energy savings
- Decision-support systems acceleration
- HW-support for sparse data

Taking it on higher level: Programming Model

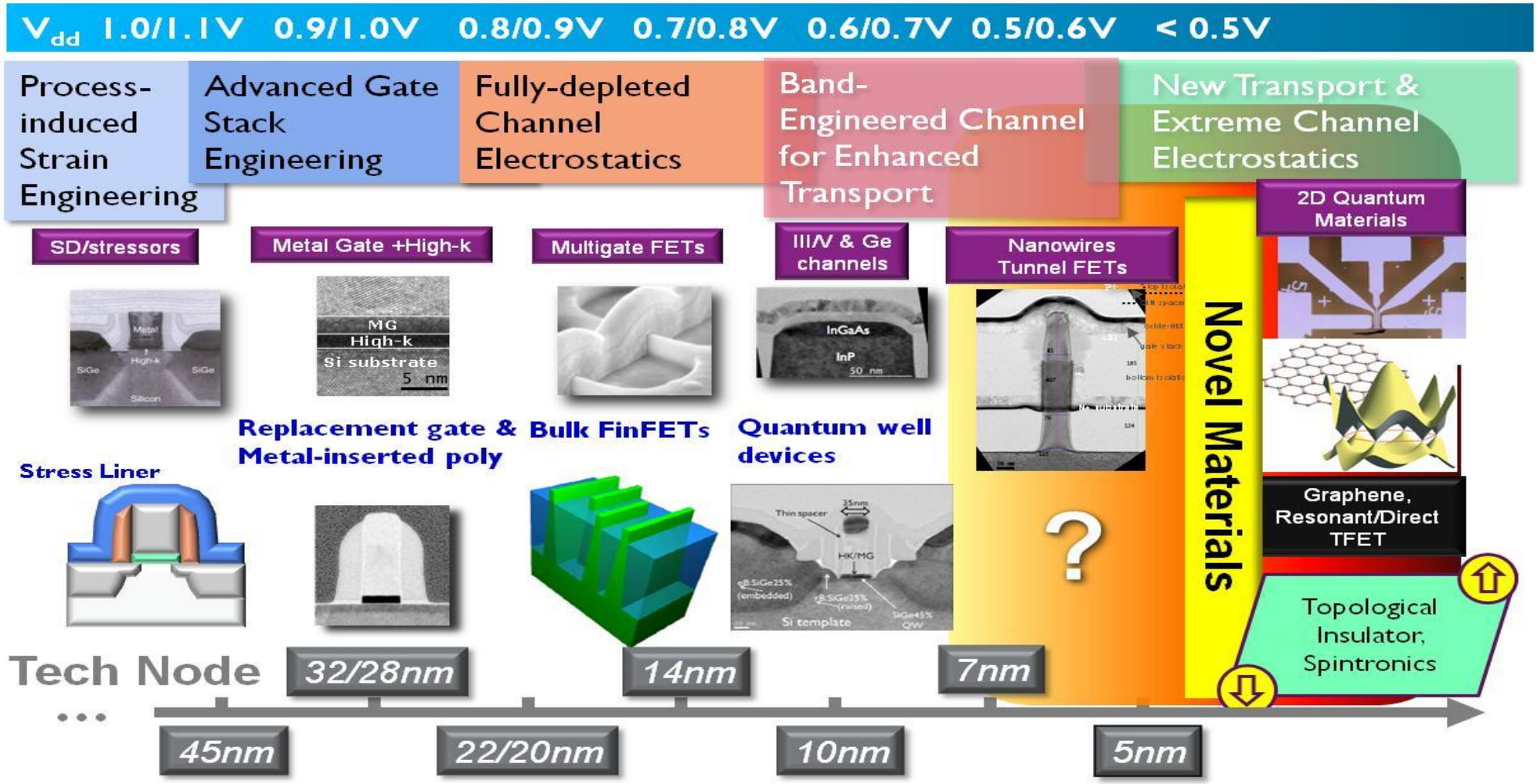
- Energy-aware programming model driving an associated ecosystem (applications, runtime and architecture) that exploits approximate and probabilistic computing and utilizes new emerging devices at the limit of CMOS scaling for radical energy savings.

Is Moore hitting a wall?



Source:IMEC

New Devices are emerging



Source: IMEC

Go below Vdd safe limits

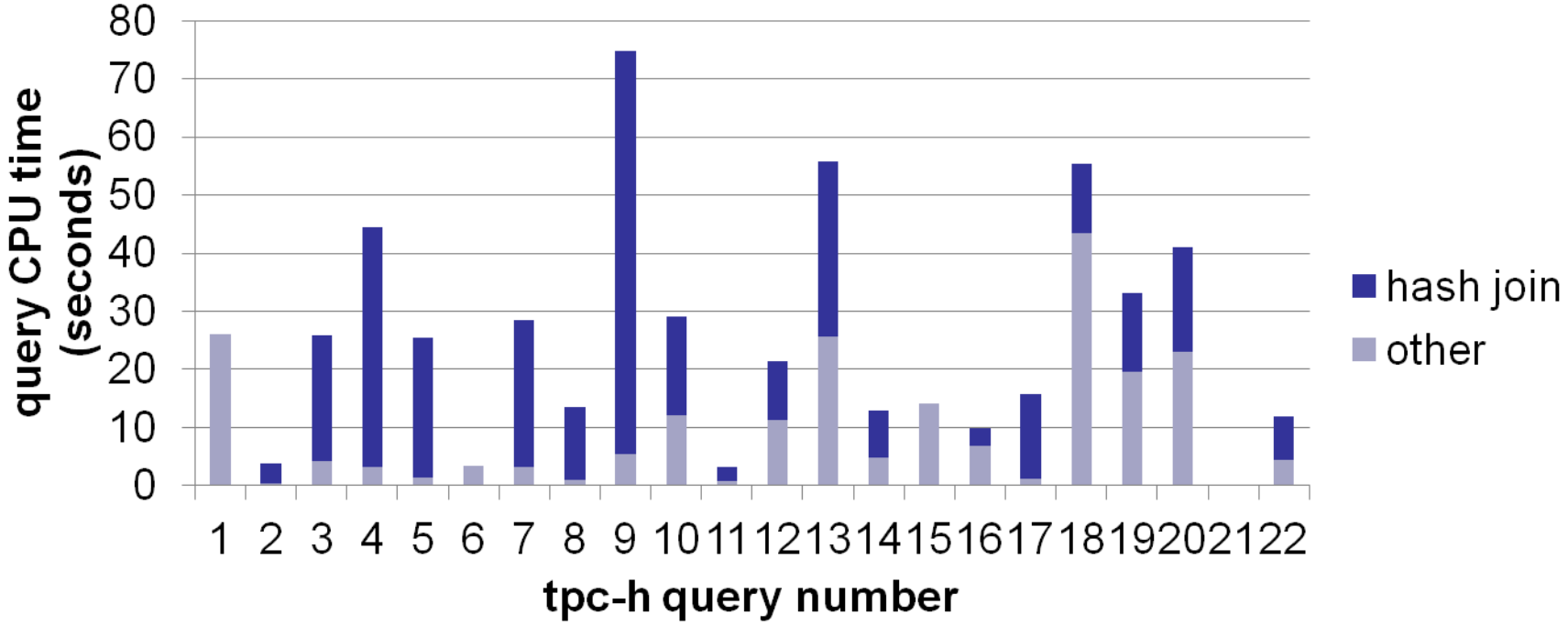
- 1.2V to 0.7V ⇨ 3X dynamic energy savings (dynamic power is proportional to $C \cdot f \cdot V^2$)
- Will have errors
  - Frequency dependent on material
- Detect and recover from errors
  - Use mix of circuit, architecture, PM to recover
- Operate at energy-efficient error-rate Vdd levels

Emerging Hardware for Big Data

- Non-volatile+3D
- ParaDIME for Programming model support for radical energy savings
- Decision-support systems acceleration
- HW-support for sparse data

Exploration

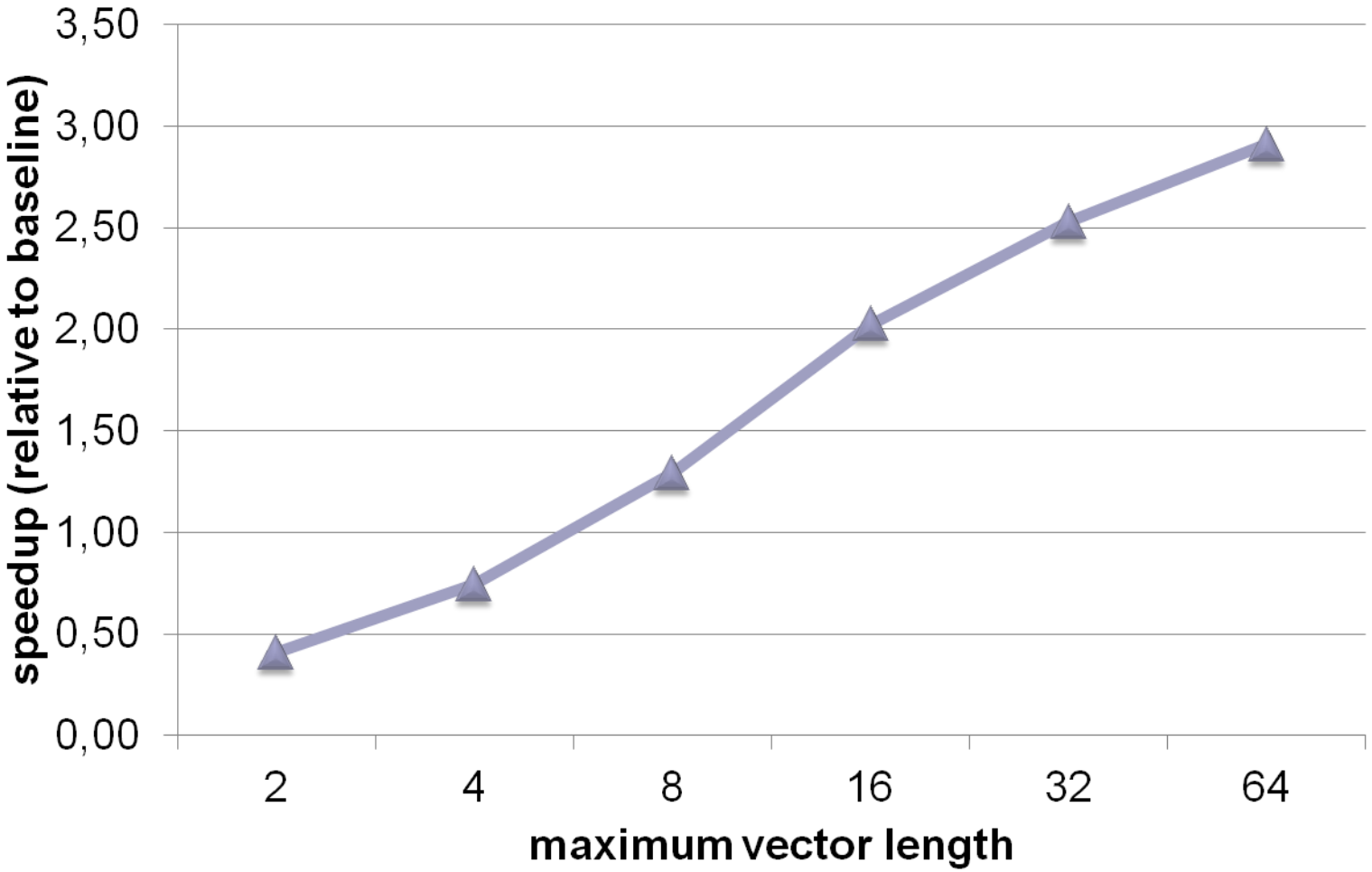
- 22 complex queries TPC-H (100GB)
- Hash joins *very* significant
- Hash tables not cache resident
- Broad-ranged memory accesses



Design

- X86-64 vector extension
- Out of Order execution
- Vector Memory Request File
  - No aliasing checks (scalable)
  - Direct access to L2
  - One request per cycle (serialised)

Varying Maximum Vector Length

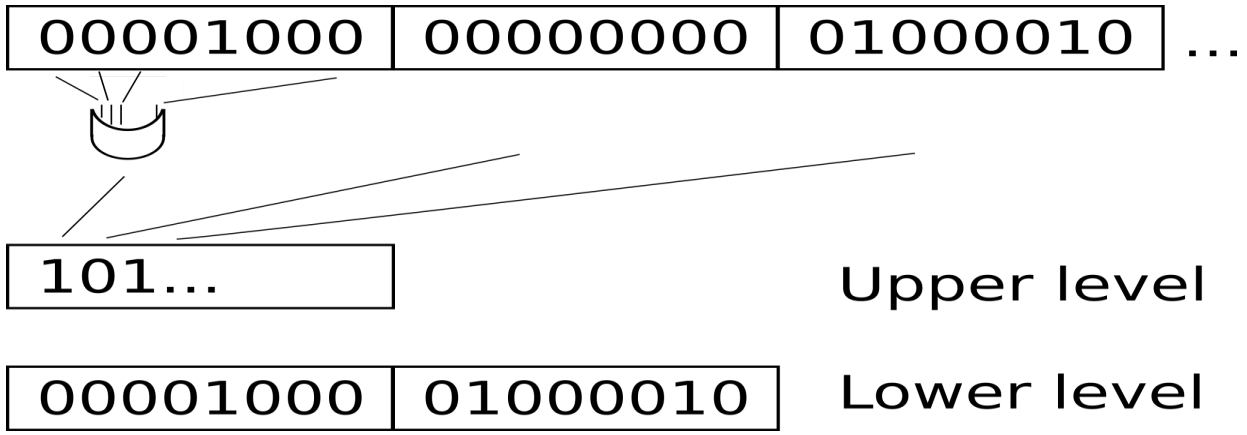


Emerging Hardware for Big Data

- Non-volatile+3D
- ParaDIME for Programming model support for radical energy savings
- Decision-support systems acceleration
- HW-support for sparse data

HW-acceleration for common big data structures: Sparse sets case

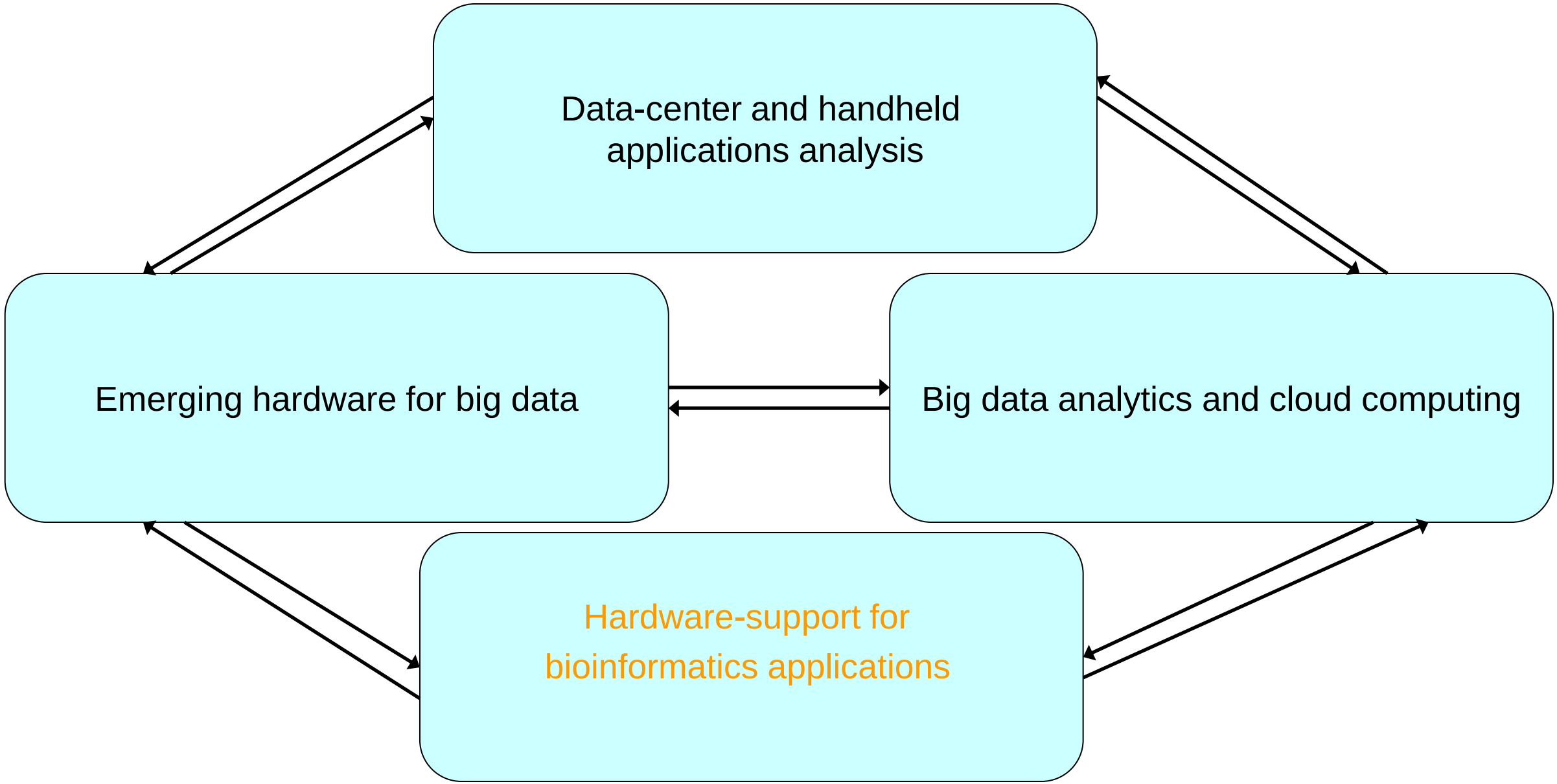
- Data mining application (ECLAT)
- Intersect operation on large sparse sets
  - Two-level bit-vectors for efficient data level parallelism (2L)
  - Indexed memory instructions, popcount, bitset
- Also initialization phase reduces intersects (from 233K to 7K)
  - Requires other instructions



Large sparse sets: Initial results

- Reduced operations in vectorized code
  - Scalar code, intersect arrays of elements
    - $9.5*10^9$ instructions
  - Vector code, intersect 2L data structure
    - $3*10^7$ instructions, $1.8*10^9$ operations
  - Assuming overhead adds up to $2.0*10^9$,
    vector version executes 4.7 times less operations
- 93% vectorized code
- Almost always use maximum vector length

Current Research Directions



HW-support for biomedicine

- Aim: Develop a special processor for personalized medicine

- In the era of personalized medicine, we envision that the way one visits a doctor could involve the following:
  - (a) one will have his/her DNA sequenced rapidly either partially or in full (if necessary multiple times in case the patient develops disease markers [Yiping2010]) so that
  - (b) the doctor can use a database of genomics, proteomics, and cytomics metadata to prescribe the optimal medicine for the patient. Creating and querying this database will require massive data mining operations over billions of DNA sequences and patient profiles. Once the treatment starts,
  - (c) proteomics will be used to closely monitor the effectiveness of the treatment and further treatment personalization could be applied.

- These three steps are currently too slow, too expensive and limited by computing bottlenecks to be viable.

HW-support for biomedicine

- Methodology: combine extensive multi-disciplinary expertise to develop
  - (i) Application-specific computer architectures (in the same vein as ANTON, whose computing cores are optimized for molecular dynamics simulation) specifically designed for bioinformatics,
  - (ii) Novel nanotechnology building blocks (such as Tunnelling FETs, Graphene and nanowires), new integration and memory technologies (such as 3D stacking, cbram, racetrack memory, sttram and memristors) offering very high integration densities, extremely high-bandwidth interconnect, and high performance, thus enabling the computational power of a conventional current supercomputer to be squeezed into single processor chips, and
  - (iii) Bioinformatics algorithms tailor-designed for optimal performance on these future hardware building blocks.

Outline

- Current Research Directions
- Past Research
- BSC Collaboration with Latin America

Past Research

- Transactional Memory
- Vector Support for Emerging Processors
- Accelerating Synchronization
- StarsS/Barrelfish

Transactional Memory

- Major focus on developing TM applications, tools, scalable HTM implementations

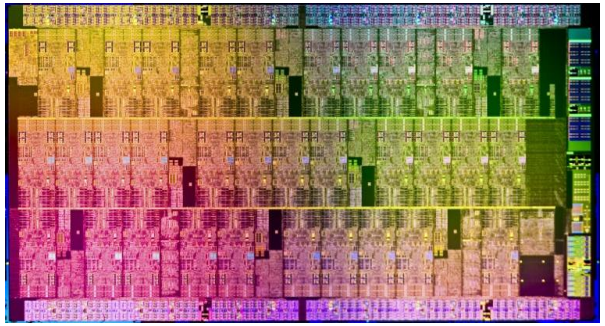
- Applications:
  - RMSTM, Atomic Quake, QuakeTM, HaskellSTMbench, Wormbench
  - C# versions of STAMP TM applications (using Bartok compiler)
  - Released publicly through www.bscmsrc.eu
  - Published in ICPE11, ICS09, ACM CF, PPoPP09
  - Lessons learned: TM not *yet* easier to program, lacks tools
  - RMSTM best paper award in ICPE11 from 110 submissions
- Tools:
  - TM debugger and bottleneck analysis best paper award in PACT2010 from 240 submissions
- HTM Implementations:
  - EazyHTM: Eager conflict detection, lazy resolution -> fast commit and aborts. Published in Micro-42
  - D1 Data cache for TM: Best paper award in GLSVLSI 2011 Conference
  - Filtering: Eliminate TM overheads by filtering thread-local data out of the read/write sets. Best paper award in HPCC09
- FPGA Emulator
  - Using BEE3 board: 4 Xilinx Virtex5-155T FPGAs, MIPS compatible cores
  - Added TLB, MMU support, cache coherence, multiprocessing facilities, implemented double ring interconnect
  - HTM implementation 16 cores per FPGA, in FCCM11, FCCM12

Past Research

- Transactional Memory
- Vector Support for Emerging Processors
- Accelerating Synchronization
- StarsS/Barrelfish

Why research vectors?

- Lots of data-level parallelism (DLP) in mobile & data-center workloads
- Vectors: efficient way to exploit that DLP
  - Compact representation of parallelism
  - Small overheads
- Recent interest on vector architectures
  - E.g. Intel's Larrabee (Knights Ferry)
- Previous vector research at UPC



What is new?

- Power and energy constrained designs
  - Vector supercomputers designed for highest performance
  - We need different microarchitectural solutions
- Traditionally not-vectorized applications
- Vector support in E2
  - Novel architecture under development at Microsoft
  - Many composable simple, power-efficient cores

Past Research

- Transactional Memory
- Vector Support for Emerging Processors
- Accelerating Synchronization
- StarsS/Barrelfish

Dynamic Filtering for Managed Language Runtimes

- Idea: Accelerate read/write barriers for managed language runtimes, use same solution for multiple problems such as STM, GC or integrity checking
- Basic idea:
  - New instruction: dyfl(addr, tag),
    - Tags are small integers distinguishing different uses
    - Have we already seen "addr" associated with "tag"?
    - If so then fall through
    - If not then branch to a lightweight trap handler (~function call)
- The filter implementation may be lossy
  - It may take extra traps
  - But it must never fall through when a trap is required

T. Harris, A. Cristal, S. Tomic, O. Unsal, "Dynamic Filtering: Multi-purpose Architecture Support for Language Runtime Systems", *ASPLOS XV, March 2010*

Past Research

- Transactional Memory
- Vector Support for Emerging Processors
- Accelerating Synchronization
- StarsS/Barrelfish

Sample Research Topic: StarSs and Barrelfish

- StarSs: BSC developed task-based programming model
  - Runtime dynamically detecting inter-task dependencies
  - Provides dataflow execution model
- Barrelfish: Microsoft/ETH developed operating system
  - Message passing based on low-level
  - Can run on shared or distributed memory
  - Designed for heterogeneous systems
- StarSs on Barrelfish
  - Leverages and combines the most attractive aspects of both: heterogeneity, message-passing, dataflow
- Is developed on the Intel 48-core Cloud Computing Chip (SCC)
  - The SCC is also message-passing based

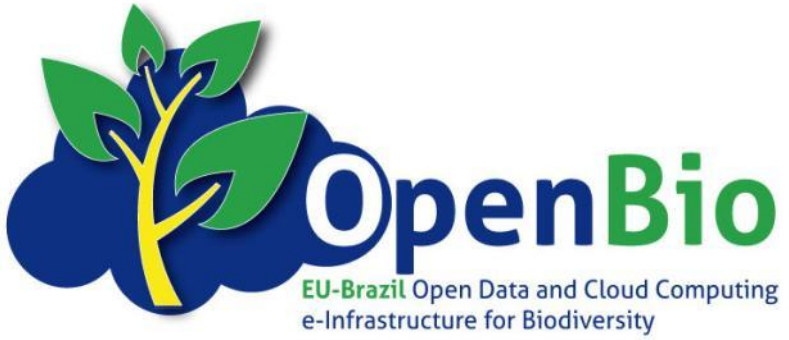
Outline

- Current Research Directions
- Past Research
- BSC Collaboration with Latin America

BSC Collaboration with Latin America

- OpenBio Project
- Risc Project

OpenBio: EUBrazil Open Data and Cloud Computing e-Infrastructure for Biodiversity



Project Aims

"Deploy an e-Infrastructure of open access resources supporting the needs of the biodiversity scientific community. This e-Infrastructure will result from the federation and integration of existing EU and Brazilian developed infrastructures and resources."

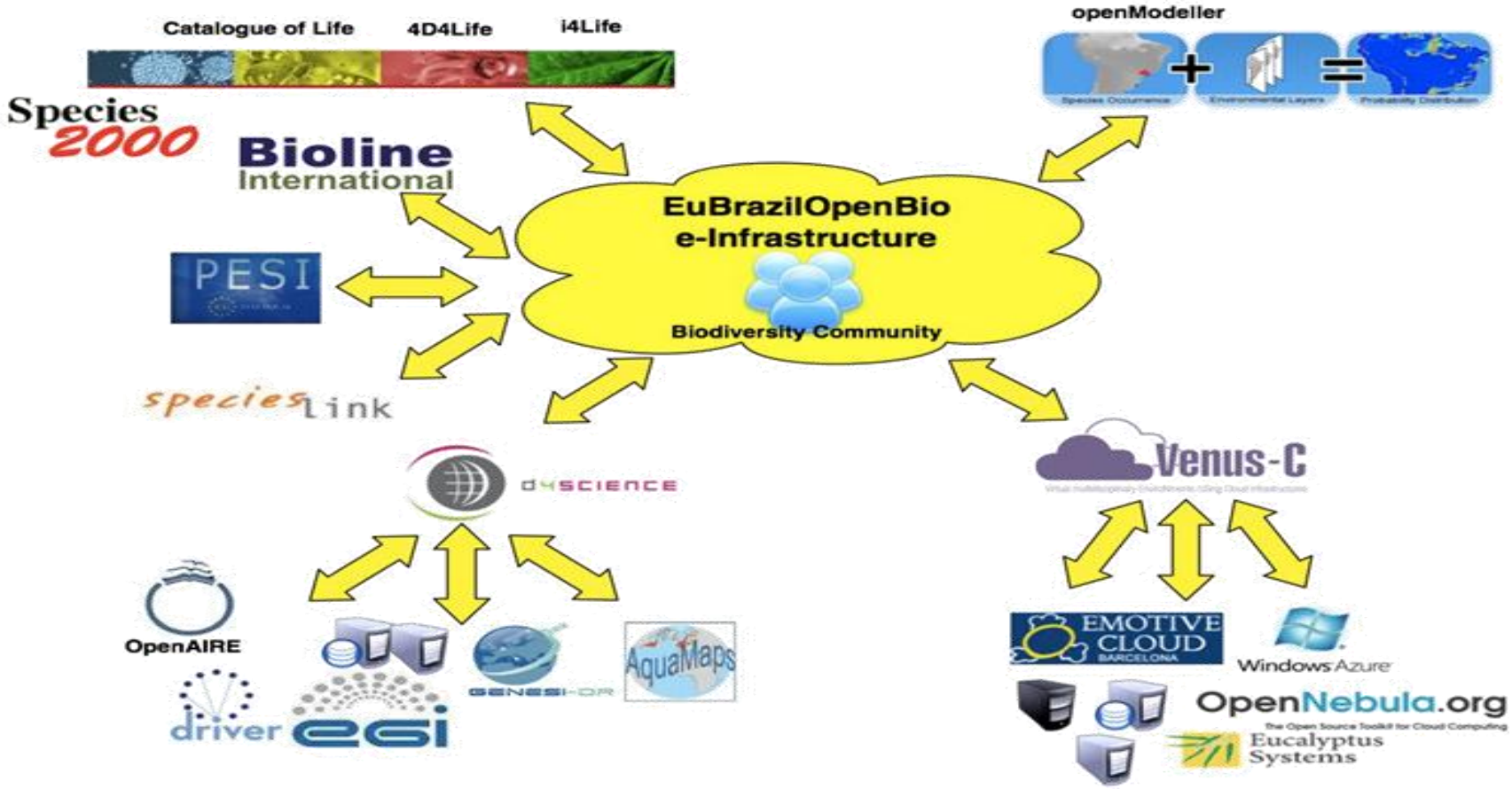
Consortium - Brazilian partners

- Centro de Referência em Informação Ambiental, SP (Brazilian Coordinator)
- Centro de Estudos e Sistemas Avançados do Recife, PE
- Universidade Federal Fluminense, RJ
- Rede Nacional de Ensino e Pesquisa, RJ (participant)

Consortium - European partners

- Barcelona Supercomputing Center – Centro Nacional de Supercomputación, Spain (European Coordinator)
- Consiglio Nazionale delle Ricerche, Italy
- Trust-IT Services Ltd, United Kingdom
- Universidad Politécnica de Valencia, Spain
- Species 2000, United Kingdom

Objectives



Objective 1: Interoperation of existing Brazilian and European e-Infrastructures in the distributed computing, scientific data and portals & platform layers

Objectives

- **Objective 2:** Provide greater focus to the integration of data software platforms running through all of infrastructures
- **Objective 3:** Identification of further future EU-Brazil collaboration in support to the biodiversity area in all types of infrastructures to demonstrate the efficiency of our approach through two Use Cases.

Expected Results and Impact

1. Open data and open access e-Infrastructure through integration of computational, storage, framework, service, and data, representing an unique value for the community that can build specialized services and workflows by simply combining these services.
2. Contribution to the federation and integration of the existing EU and Brazilian developed infrastructures and resources through Catalogue of Life, openModeller, D4Science-II and Venus-C.
3. Development of two Use Cases: Integration between Regional & Global Taxonomies and Data Usability and the use of ecological niche modelling Enhance
4. Enhance EU-Brazil collaboration in biodiversity area through policy dialogue and Joint Action Plan.

BSC Collaboration with Latin America

- OpenBio Project
- Risc Project

RISC

A Network for Supporting the Coordination of Supercomputing Research between Europe and Latin America

http://www.risc-project.eu

Objectives

- Deepening strategic R&D cooperation between Europe and Latin America in the field of High Performance Computing (HPC) by building a multinational and multi-stakeholder community through capacity building, awareness building, networking and training events;
- Identifying strategic research clusters in order to foster targeted research collaboration in these areas;
- Assessing the ICT collaboration potential for the two regions – EU and Latin America;
- Producing a Green Paper on High Performance Computing and Supercomputing Drivers and Needs in Latin America;
- Producing a Roadmap for High Performance Computing and Supercomputing strategic R&D in Latin America;
- Enhancing HPC R&D policy dialogue between policy makers and stakeholders from EU and Latin American HPC communities;

Impact Research Areas

- **HPC and Supercomputing - driver for Innovation** (Innovation and HPC, *etc*)
- **Computational Biology** (such as: Advanced Modeling of the Genome, Genome Sequencing, Modeling of Epidemics, *etc*)
- **Oil Exploration Advanced Modeling** (such as: Advanced Computational Methods and Techniques for Oil Exploration, *etc*)
- **Natural Disasters Modeling and Simulation** (such as: Hurricane a Catastrophe Modelling, Air Pollution Modelling, *etc*)
- **HPC and Supercomputing as a platform for research in the Industry and Academia** ;

Expected Results

- Green Paper on High Performance Computing and Supercomputing Drivers and Needs in Latin America
- Roadmap of High Performance Computing and Supercomputing strategic R&D in Latin America
- Strategic research clusters established
- Fully functioning network focusing on activities to support and to promote coordination of the HPC and Supercomputing research between Europe and Latin America

Next Events

- Winter/Summer School Buenos Aires 25th July 2012
- Advanced Workshop Colombia 2013
- Summer School Rio de Janeiro 2013
- Advanced Workshop/ Awareness Raising Spain 2013

Partners

UNIVERSIDADE DE COIMBRA

Universidad Veracruzana

The RISC project is co-funded by the European Commission's 7th Framework Programme (Grant Agreement no: 288883)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Academic Liaison with Mexico

- **HUMAN CAPITAL Agreements**

- CONACyT- BSC Agreement (2012)
  - 5 postdoctoral x 2 years/each
  - 10 PhD x (3 + 1) years/each

- **Research Agreements**

- CINVESTAV – BSC
- CIC-IPN --BSC
- U. Veracruzana- BSC
- U. De Guanajuato - BSC

Further Mexico collaboration

www.bscmsrc.eu
Thank you!

Questions?