Introducing Cloud Computing into STEM Curriculum Using Microsoft Azure

B. Ramamurthy

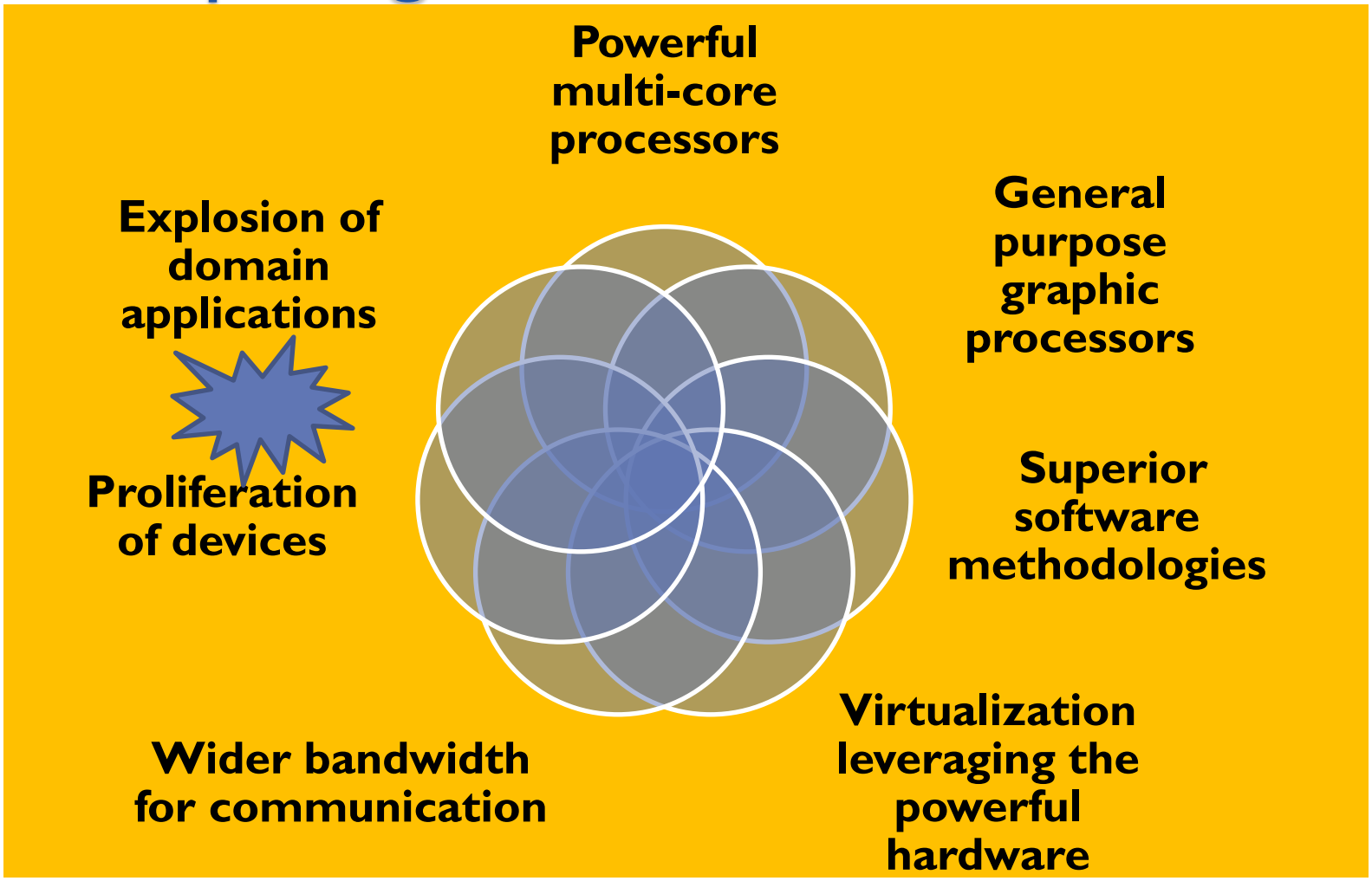
Computer Science and Engineering Dept.

University at Buffalo

bina@buffalo.edu

http://www.cse.buffalo.edu/faculty/bina

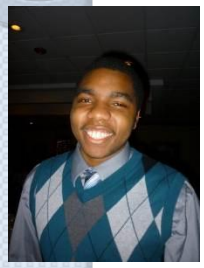
Introduction: A Golden Era in Computing

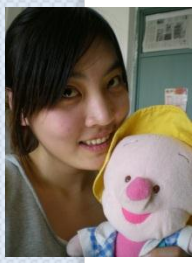


Topics for Discussion

- People involved in this project
- Challenges in introducing Cloud Computing
  - Educational imperative
  - Efficacy of existing methods
  - Cloud models
- General approach introducing cloud computing to STEM curriculum
- A Transformative Pedagogy
- Demo 1: Name2Face : Addressing CSE audience
- Demo 2: Digital Biology: Addressing broader STEM audience
- Features of Azure and Challenges
- Summary

People on this project





- This is a work in progress.
- Damian: an ***undergraduate*** sophomore honors student who chose to work on this project as his ***summer research***
- Suchismit: a ***graduate*** student with deep theoretical background who wants work with cloud computing ***research***
- Ying: an enthusiastic ***female graduate*** student who likes to implement ***projects on the cloud***

Challenges in introducing Cloud Computing

- More people are familiar with cloud computing when compared to the Internet during its infancy
- Educators as well as students want to learn and use the "cloud" and "cloud computing"
- However,
- CSE and STEM (Science, Technology, Engineering and Mathematics) curriculum is quite rigid steeped in tradition and standards
- Current curriculum is so overloaded there is very little opportunity for new material
- Server-side and large-scale application development is indeed a challenge with ***storage needs***, ports, web servers, firewall and ***vulnerabilities***
- Educators are overwhelmed preparing for the dense content of the existing curriculum, newer concepts take a backstage

Educational Imperative

- Need a highly effective learning approach/pedagogy for introducing emerging concepts to undergraduates
- Concepts: parallelism, algorithms for knowledge discovery, services-oriented design, large-scale data-intensive analytics, design for high-performance and automatic load balancing, large-scale storage and monitoring for access control and performance.
- ***Cloud computing*** encompasses these concepts.
- Need to provide broad access to these concepts irrespective of the discipline
- Need to excite and ***engage the net-generation***.
- The learning should be ***organic***.

Efficacy of existing methods

- Cloud computing discussed as a part of distributed systems course.
  - Ex: I was awarded an NSF CCLI grant for data-intensive computing certificate program.
  - But this is just one program, we need a ***systemic change*** and a solution that can be adopted and adapted widely
- Microsoft and other organization's effort during conferences such as SIGCSE to promote emerging technologies
  - Excellent informational overview; however typically not aligned with the curricular needs and limitations
- …

Cloud Models

- Amazon EC2: nuts and bolt: good for experienced students (IaaS)
- Google App Engine: software-based; needs knowledge of services-based programming (SaaS)
- SalesForce.com; polished applications; more for business-oriented use (AaaS)
- Microsoft Azure Platform: (PaaS)
  - Provides a logical framework for modeling and design, and for teaching important computational concepts
  - Offers easy migration into cloud for schools who are already using MS academic licensing (Visual Studio)
  - Provides easy to understand building blocks: web role, worker role, blob storage etc.

Our Approach

- We will develop a complete set of self-contained modules ***customized and aligned*** to the contents of specific courses.
- ***Sustainability*** through the levels of a curriculum allowing for a spiral mode of learning.
- Educators can use these readily usable modules to introduce Azure cloud computing seamlessly by ***spending about a week's lecture*** (about 3 lecture hours).
- Modules will be updated with versions of Azure.
- We will use an application ***Name2Face*** as a core-seed application and extend it incrementally to address issues across curriculum.
- It is a simple application that is more a metaphor for mapping, correlation and association ***engine*** that powers most modern applications.

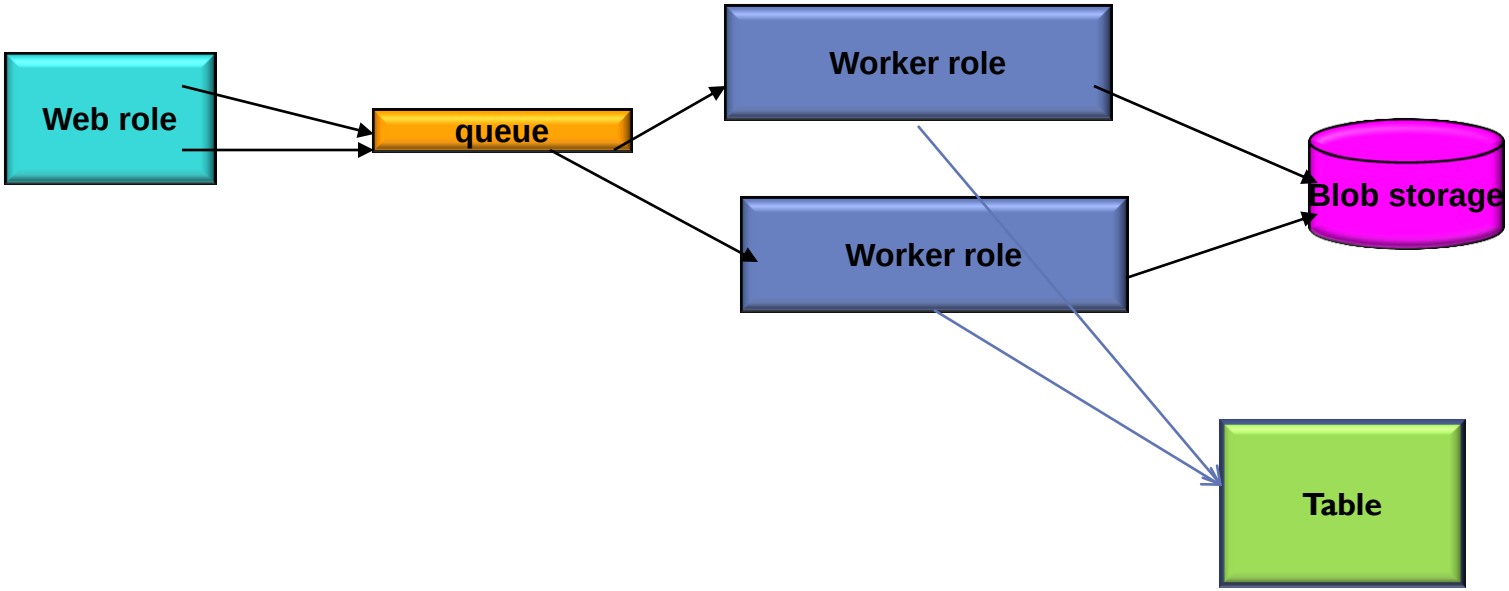
Name2Face Details

- We have chosen an easy to understand application.
- Data Structure(s): A list of faces and names.
- Algorithm: Given a name, the application displays the images of faces associated (tagged) with the name.
- Demo 1: Here are more Azure details and the demo of the application.

More details

- We use web role, worker role, and blob storage
- Web role presents the user interface to load the image, tag the image, and to search for an image given a tag/name
- Web role talks to worker through message passing via a queue
- Message contains the tags, path to the image
- Worker role creates a thumb nail of the image (representative of any work) and loads into the blob storage
- Illustrates: loose coupling, message passing, blob storage service, service authentication through PKI
- Worker roles can be complex and implement any algorithm being taught in a CSE course.

Application Architecture for the demos



Transformative Pedagogy

- Includes a set of modules that provide ***entry points*** into cloud computing for educators
- These are ready to use, self-contained modules ***aligned to the contents*** of the courses at various levels
- These modules allow for ***repeated exposure*** to fully comprehend the broad utility of the cloud and to ***gain competence*** in using cloud computing
- ***Server-side enterprise-level*** application development is a simple extension of monolithic single-tier development
- Next we present a few ***representative modules*** of this pedagogy
- Detailed report of this pedagogy has been sent to Microsoft and will be available in a future publication.

Module 1: Data Structures

- CS1and CS2 are the first courses in Computer Science;
  - Azure cloud installation and the Name2Face are introduced with various ***data structures and algorithms***.
  - Develop and ***deploy on Azure cloud***.
  - In memory data-structures.
  - Some representative outcomes:
    - Students exercises can compare performance of arrays, list, maps, trees etc..
    - Emphasis on worker role
    - Development and production environments provide the students with professional practices in software development.
    - This module well suited for K-12 as well as Community Colleges.

Module 2: Algorithms

- This covers the ***programming language*** course and the advanced ***algorithms*** course (at 200 and 300 levels).
- Azure cloud's Blob storage is introduced for large scale storage for Name2Face.
- Some representative outcomes:
  - Student exercises can study the various sort, search algorithms.
  - Loose coupling; ***enterprise level development***
  - Web role and worker role can be utilized creatively to illustrate programming language concepts; Lambda expression, first class objects, parallelism constructs

Module 3: Database and Web services

- Distributed systems, database systems with 4XX course levels
- Persistence models, ***data models, cloud database and database as a service***.
- Learning outcomes:
  - Flexibility of the Azure Table structure
  - Working with SQL Azure data models
  - REST web services access to storage
  - Loose coupling web and worker roles and the storage
  - Scalability of tables and blobs; load balancing, multiple instances of roles
  - Lesson in realizing security through certificates

Module 4: Applications

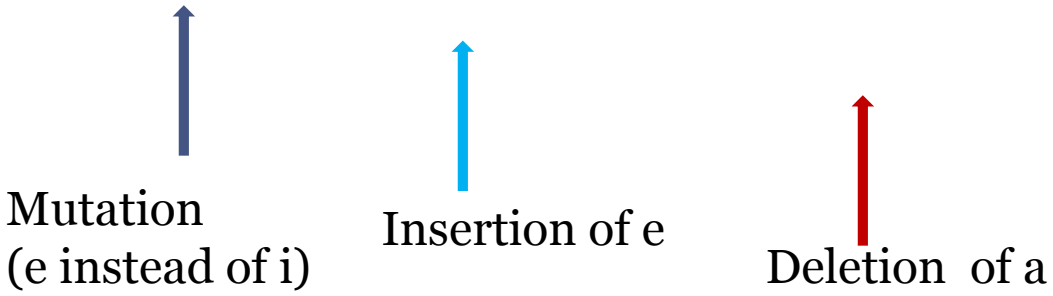
- STEM application domains: data-intensive computing, digital biology for example; ***Capstone projects***
- Applications drive the advances in technology
- Expected outcomes:
  - Introducing Azure cloud to other STEM disciplines
  - Provide entry point for non-CSE students
  - Service bus, access control, caching (for CSE)

Digital Biology

- 20th century is a century of physical sciences; 21st century is a century of great understanding of life
- Biological infrastructures
- Data-intensive computing for biological problems: specifically genetics, human genome exploration, analysis of high-throughput sequence data, etc.
- We will discuss some background information and demo a biological index/search application on Azure.

Unstructured data example (English language)

- Thes booke is on the tble
- Thes booke is on the tble



- This book is on the table

Vertebrate Mammalian Genome

Size 247259409 bytes : 247M (one of many files)

TATTAAAAGCACATCTGCCCCAAGAGGCCTTCCCAGACTAAACGCCACTTTTCCTCATTTCCCATTCCCTTCTGCGTCAC
CCTAACTTTCTTTTCCCCCGCTTTCCAGCCTCACAGCACTTATGTTTATATTTGTAATGTTATTTATTTATATTGATCTC
TGTCTCCCCGACCCCTAGACTGTGAGCTCACTGTGGGCAGGGATTGTCACTGTTTACTGTTGTATTGTGCTTTCCCAGGG
GCTTCATACAGTGCTTTGCACACAGTAAGCACTTAAATGCGGTTGGATGAATGAATGAATTTACTACAGCACAGTAGCTG
AGAACACAGGGAGACACTGTGATGTGAGGAGGGCTGTATCTGACTGTATTGTACCCACCACAGTTCTTGGCATAGTGTGT
AGCACACAGCAAGAACTTATATGTCCCAATTATTTTATTATTTTTTATTATCTGGAATTAGCAATCTACCACTTTGCCCT
AGATTGTGAGCCCCACCTGTGTCTTGTATCTACCCTAACGTTTAGTACAGTGCTTTGTATACAGTAAGCACCTAAATACC
CCAATTATTATGCATTGTGATGTTCTCGGAATGCTACTAGTCATTTGGATTTTTTTCCATGGGATGATTTAATACTTTCA
AGCATTATTCATTGATCTTTCTGGGTTCTAACTGGGTGGCAAGTATAATCAGTCTACTTCAAAGATGGAGAGAACATCCT
AGTGTCTGGCAAAGAGGTGGCTAGATCTTAAAGGGACAGGGGAACGCTAGCAAAAAGACTCTAACCCCGATGTTATCTCC
TGATCTCCACGCCGCCTCTCTTAGAAGGCCTGCACTGTTTCCTTGCACTATTAATAACAAGGCACTAGCAAAGAAGACTC
CCCTCGGAAAAGGATCTCTGAACAACTATTTTGAAAGTTATTTCAAAAGAGTTTTCTATAAGCAATGCATTGAAAATATA
AATGCACTTTAAACATTAAAAACTCCATGGGAAAAAAGTTAGAGTATAAACATACATCAAAATATGCCAAAATAGAACAC
ACAGTGTTTATCTGAAGATGTAGCTGTGACACCTAAAATAAAGCATATACGCTGAAAGGATGATATTAAACATAATGGCC
TGGATTAGGCCCATAAAACCAATCAAGTTGAAAGGGAATGGCAAAATATGTTGTATAATTTTTTTACATTAAAAGCAAGA
TTTCTTGAATTTATCACTCCAACAGATAAAACTTTACTTCATCAAAATGACTTGCATAATGCAGTAGAATAAATTTCATG
ACATATGGATTTTATAAAAGAAAGGGAAAAAGAATGTGCATTCACAGGAGACTCATCATACGGAAGACCTATATTTTTTA
AGGTTCTGTGGTTCATTTTTAAGAACGAAACTTAATCTTTAACTCGACCCGCTCTTAAAAATTGTGGTTCATGTGTAAGT
TAAAAAGATCAGTGGGTTGAGTTTTCTGGCTGGCAGTTAATCATCTGGAAAAGCTCACATTTATCTGGAACAGAGTAAAT…..

Reference sequence and query sequence

- Assume the reference sequence given in the last slide:
- Below is a six residue peptide for DPZ, a protein which acts as a tumor suppressor. HPV (human papillomavirus) is a virus that binds to DPZ and degrades it, causing oncogenesis.
  - Codon sequence for: Glycine – Leucine – Glycine – Phenylalanine– Asparagine– Isoleucine
  -
  - GGT, CTT, GGT, TTT, AAT, ATT
- Query sequence: GCTCTTGCTTTTAATATT
- Now add all the possible mutations, insertions, deletions in the reference sequence ☹
- Diagnose using (data-intensive) computational methods, the probability of oncogenesis in the person with a given reference sequence.

Microsoft Biology Initiative (MBI)

- Microsoft Biology Foundation (MBF)
- Microsoft Biology Tools (MBT)
- Excellent resource with many packages with Apache license.
- From Microsoft Cambridge Research.
- Fertile source for applications that can be staged on Azure.
- This leads us to demo 2.

Features and Challenges

- Great features:
  - Access to cloud computing through Visual Studio
  - PKI-based authentication for storage/server
  - Table storage is very intuitive and highly flexible
- Challenges
  - The message queue between the web and the worker roles is a bottle neck
  - The interaction of this queue under .Net's parallelism features need to be further studied by Azure designers.
  - Internal workflow and internal operation of Azure to be clarified: what can and cannot the web role do? Worker role? Best practices?

Summary

- Cloud is an exciting and high impact infrastructure
- MS Azure offers excellent resources for teaching cloud computing at all levels
- It is imperative that our workforce is instructed on this emerging technology to enable them to create ***innovative cloud-based computing models and novel research directions*** to solving urgent problems
- In this talk we demonstrated a systematic approach that provides ***entry points*** and pathways to access Azure cloud computing.