Modern Science & Society Transformed by Compute & Data

- New Paradigms & Practices
  - End-to-end: Seamless access, aggregation, interactions
  - Integrative, multi-scale, online
  - Data-driven, Data/Compute-intensive

- Multi-disciplinary/scale collaborations
  - Individuals, groups, teams, communities, networks

- Unprecedented opportunities, challenges
Many Challenges

• Computing
  – Multicore; large and increasing core counts, deep memory hierarchies
  – New prgm. model, concerns (fault tolerance, energy, etc)
  – New models & technologies: Clouds, grids, hybrid manycore, accelerators, deep storage hierarchies, ...

• Data
  – Generating more data than in all of human history: preserve, mine, share?
  – How do we create “data scientists/engineers”?

• Software
  – Complex applications on coupled compute-data-networked environments, tools needed
  – Modern apps: 10^6+ lines, many groups contribute, take decades

• People
  – Multidisciplinary expertise essential!
    • Appropriate academic program, career tracks...

Advanced Computing Infrastructure

• Large scale, distributed, heterogeneous, multicore/manycore, accelerators, deep storage hierarchies, experimental systems

Titan - Cray XK7
- 27 PF / 56 K cores
- 16-core CPU + GPU
- Gemini 3D torus
- 710 TB memory

Sequoia – IBM BG/Q
- 20 PF / 1.6 M cores
- 18-core processor
- 5D torus
- 1.5PB memory

Worldwide LHC Computing Grid
- >140 sites;
- ~250k cores;
- ~100 PB disk

XSEDE
- Worlds Largest Grid
- 11 Resource Providers

Modern Datacenters
- 1M servers
- 50-100 MW

Special Purpose HW (Anton)
- > 100 time acceleration of MD simulations
Data Crisis: Information Big Bang

"Data generation == 4 x Moore's Law"

There is a pending crisis in archiving… we have to create long-term methods for preserving information, for making it available for analysis in the future." 80% respondents: >50 yrs; 68% > 100 yrs
Big Data in Science & Engineering:
Modern network/instruments/experiments/…

Large Hadron Collider

Big Data Scale
- Peta
- Tera
- Giga
- Mega
- Kilo

Data at Rest
- Traditional Data warehouse & Business Intelligence

Data in Motion
- Complex Analytics

- Biased (Un)structured

Up to 10,000 times larger

Up to 10,000 times faster

Decision Frequency
- yr
- mo
- wk
- day
- hr
- min
- sec
- ms
- μs
- Real-time

Ack. A Blatecky

Image credit: Roger Smith/NOAO/AURA/NSF

Image credit: Valerio Mezzanotti for The New York Times

Image credit: Roger Smith/NOAO/AURA/NSF

Above is proposed image
Exploding Data Volumes in Lifescience

- Economic advantages depend on the data you have and your ability to transform that data into meaningful and timely insights
  - Industries nimble enough to interpret & use the data in new ways to add value will be the leaders
- Traditional decision-making structures must be adapted to incorporate data scientists in business and research
  - Integrate compute and data into reasoning/problem solving processes
- Need to move away from our fixation on data size - data quality & the ability to analyze it are more important

- Working harder is not enough .... We need to work smarter!
Rutgers Discovery Informatics Institute: RDI²

- Fundamentally integrated research, education, ACI and industry partnerships to address core CDS&E / BigData challenges
- Broaden access to state-of-the-art computing technology; integrate multidisciplinary research with ACI and industry partnerships
- Enable large-scale data analytics, computational modeling, and simulations, all of which are playing an increasingly important role in both academic and commercial research and innovation.
- Only university-based advanced computation center in NJ, and one of about ten in US, with an industry partnership program

Rutgers University, IBM Open Supercomputer Center

Today’s machines are taking on challenges ranging from cancer research to “Jeopardy” champions.

Rutgers, IBM to Build HPC Center focused on Big Data Analytics

Supercomputer to Power Industry, Rutgers Research

IBM BlueGene supercomputer to help New Jersey industry

Rutgers Teams With IBM to Build Powerful High-Performance Computing Center in New Jersey

Rutgers puts a supercomputer out of its own making to address big data challenges
Key Programmatic Areas

Research
- Provide Rutgers researchers access to computational resources and technical expertise necessary to increase accuracy and scale of their research
- Promote interdisciplinary collaborations to increase grant competitiveness

Advanced Computing Infrastructure
- Deploy a balanced cyber infrastructure composed of advanced data and compute and communication capabilities
- Experimental platforms (e.g., accelerators, novel storage and network technologies)
- Expertise (system and application facing)

Education and Training
- Variety of education and training programs for faculty, students and industry
- Masters degrees, certificates, technical modules, industry-specific workshops

Industry Engagement and Economic Development
- RDI2's Industry Partnership Program will assist private firms in overcoming the cost and knowledge barriers associated with advanced computation
- RDI2 will promote economic development by attracting new firms to New Jersey and encouraging existing firms to stay in-state

RDI2 Advanced Computing Infrastructure

The goal of the RDI@ ACI is to deploy a balanced cyberinfrastructure composed of advanced data and compute and communication capabilities that can drive research, innovation, and economic development

- Phase I – “Excalibur” IBM Blue Gene®/P Supercomputer
  - 2,000 Nodes, 8,000 Cores
  - 24 Terabytes of RAM
  - 300 – 400 Terabytes of memory
- Result of an Rutgers-IBM partnerships

- Phase II Plans:
  - Acquisition of a 10 – 12 petabyte storage container with co-located analytics
  - Upgrade to Blue Gene/Q (or equivalent)
  - Connectivity to national cyberinfrastructure

- Phase III Plans:
  - Deploy one of the top academic supercomputers in the world
RDI²’s Industry Partnership Programs

**NSF Cloud and Autonomic Computing (CAC) Center**
**NSF Center for Dynamic Data Analytics (CDDA)**

CAC/CDDA are multidisciplinary NSF centers of excellence in cloud and autonomic computing research

Foster long-term collaborative partnerships among industry, academia, and government

Have a well-established Industry Partnership Program and many industry partners

---

RDI²’s Industry Membership Program

- Tiered membership structure provides industry members with a range of benefits and resources including, but are not limited to:
  - Access to leading experts in ACI and its applications
  - Access to limited-access technology reports
  - Access to top graduate students
  - First right to negotiate commercial licenses on certain technologies
  - Invitation to participate in bi-annual center-wide meetings for program evaluation
### RDI² Tiered Membership Model

<table>
<thead>
<tr>
<th>Membership</th>
<th>Level 1 Advisory Board Member</th>
<th>Level 2 Advisory Board Member</th>
<th>Level 3 Associate</th>
<th>Level 4 Associate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steering Committee Member</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Advisory Board Member</td>
<td>Yes – 2 seats</td>
<td>Yes – 1 seat</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Biannual Meetings</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>First right to negotiate commercial licenses</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes, but only in industry thrust</td>
</tr>
<tr>
<td>Royalty-free access to institute technology</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes, but only in industry thrust</td>
</tr>
<tr>
<td>Access to ACI Experts</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Membership Fee</td>
<td>100K/yr cash only</td>
<td>50K/yr Cash/In-kind</td>
<td>30K/yr Cash/In-kind</td>
<td>15K/yr Cash/In-kind</td>
</tr>
</tbody>
</table>

### Membership: A Pathway to Further Collaboration

- **RDI² Industry Membership**
  - Collaborative Research Agreements
  - Sponsored Research Agreements
  - Service Agreements
  - Student Recruitment Opportunities
Summary

• Large scale compute and data are revolutionary science, engineering and society
  – Advanced Computing Infrastructure + Big Data
  – New opportunities and challenges
  – Requires new thinking in research and education practices
  – Multidisciplinary collaborations are critical!

• Rutgers Discovery Informatics Institute (RDI²)
  • Integrating multidisciplinary research, ACI and education @ Rutgers – http://rdi2.rutgers.edu/
  – Industry Partnership Programs
    • NSF Cloud and Autonomic Computing Center (CAC) – http://nsfcac.rutgers.edu/

• Some upcoming event (Spring 2013)
  – NJ Big Data Alliance
  – NJ Virtual Computational Institute

Thank You!

Manish Parashar, Ph.D.
Prof., Dept. of Electrical & Computer Engr.
Rutgers Discovery Informatics Institute (RDI²)
Cloud & Autonomic Computing Center (CAC)
Rutgers, The State University of New Jersey

Email: parashar@rutgers.edu
WWW: http://rdi2.rutgers.edu / http://parashar.rutgers.edu