

Performance vs. Accuracy Trade-offs for Large-scale Image Analysis Applications

Vijay S Kumar, Tahsin Kurc, Jun Kong,
Umit Catalyurek, Metin Gurcan,
Joel Saltz

**Department of Biomedical Informatics
The Ohio State University**

IEEE Cluster 2007
September 18, 2007

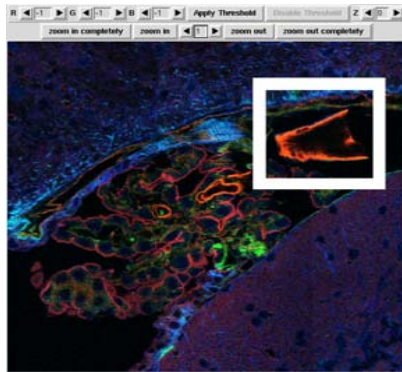
Organization of the talk

- Motivation: Digital microscopy, pathology
 - Image analysis for cancer prognosis studies
- Adaptive data analysis
- Basic image-analysis algorithm

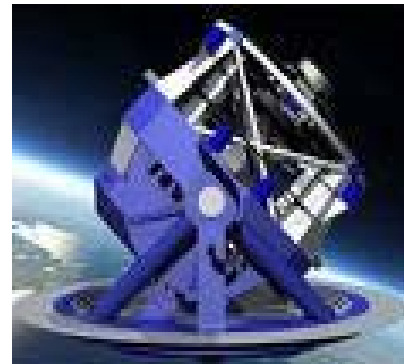
- Adaptive processing heuristics
- Parallel execution on cluster
- Performance Results
- Conclusions

Motivating application scenarios

Pathology



Astronomy



Large data volumes, High data rates
Complex data processing operations

- Digital Microscopy: 500 MB/min
- 2D images acquired at upto 150x magnification (*60 Gigabytes per slice uncompressed*)
- **Whole-slide image analysis: High turnaround times**

- Large survey telescopes: 12 TB/night
- Science requirements on performance are **stringent**
- **Near-real time data analysis**

Adaptive applications

- Applications can execute at different performance levels
 - Higher performance level → Lower accuracy of output
- Trade-off controlled by **application-level parameters**
- Application-level Quality-of-service (*QoS*) requirements:
 - *Maximize average accuracy across all tasks within a deadline*
 - *Given a minimum average accuracy per task, minimize time*

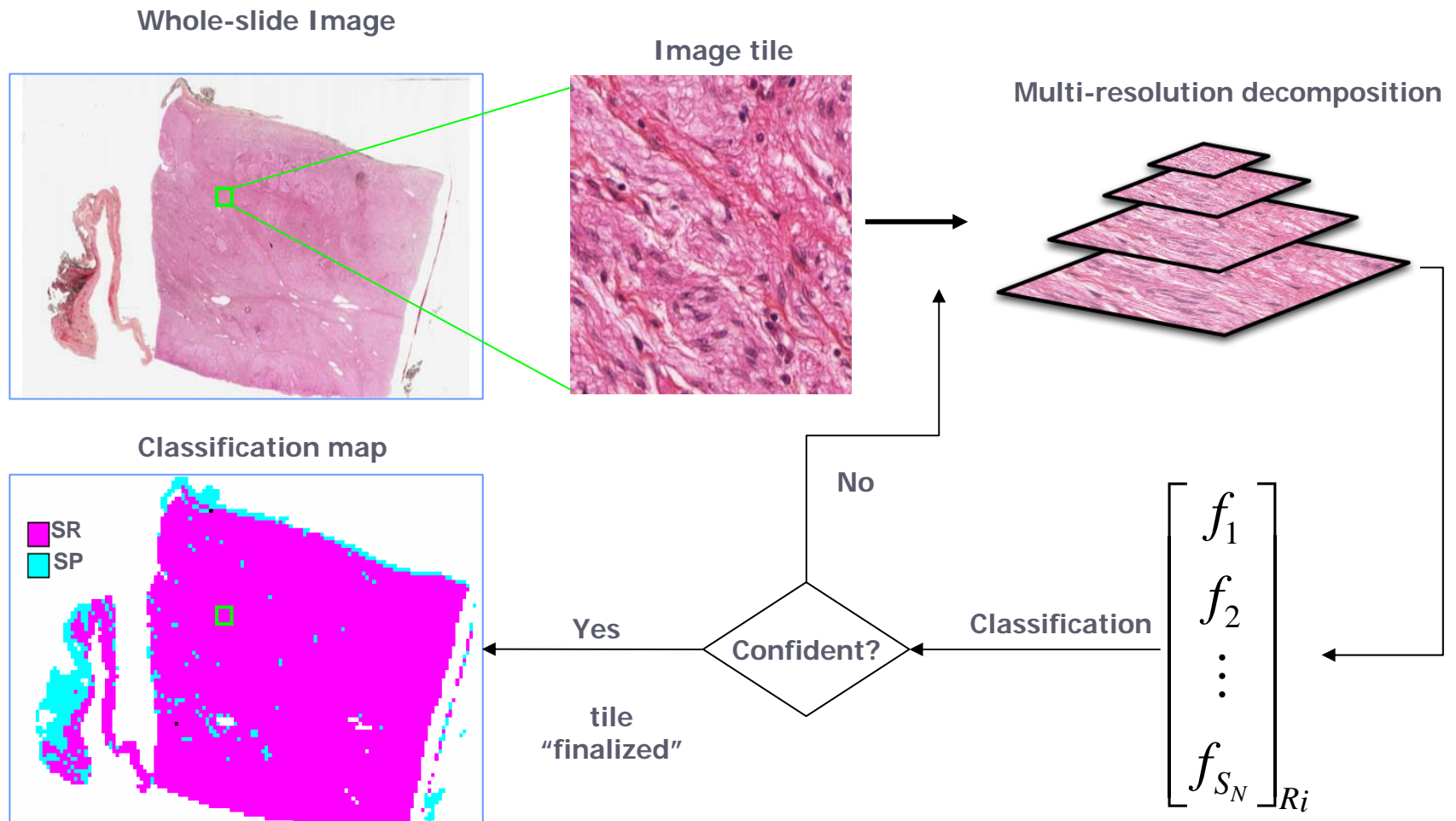
Resolution	Processing time	Accuracy
<i>256 x 256</i>	<i>1088 s</i>	<i>0.509</i>
<i>128 x 128</i>	<i>483 s</i>	<i>0.483</i>
<i>64 x 64</i>	<i>374 s</i>	<i>0.446</i>

Adaptive multi-resolution strategy
used for image analysis

GOAL: Framework to support QoS in
large-scale image analysis

- Application: Express performance parameters, QoS requirements
- System: Estimate accuracy vs. performance characteristics, Adaptive processing (map requirements to parameter values)

Adaptive multi-resolution strategy



Computer-aided classification of neuroblastoma

Problem definition

- To process tiles in an image such that user QoS requirements are met.
- Maps to a scheduling problem:

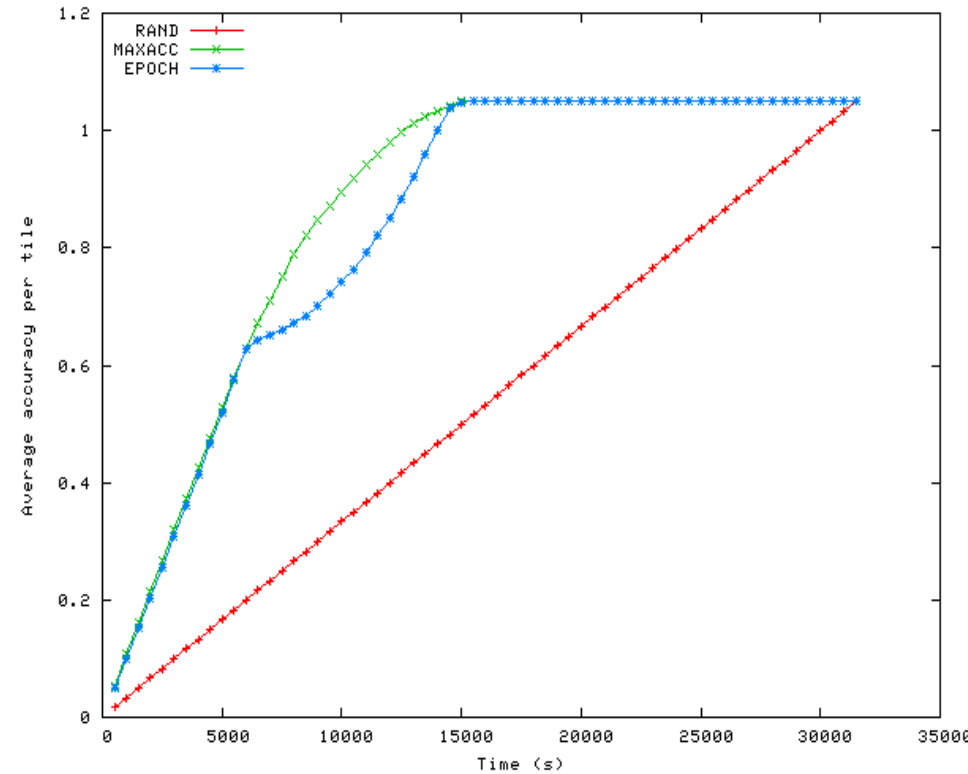
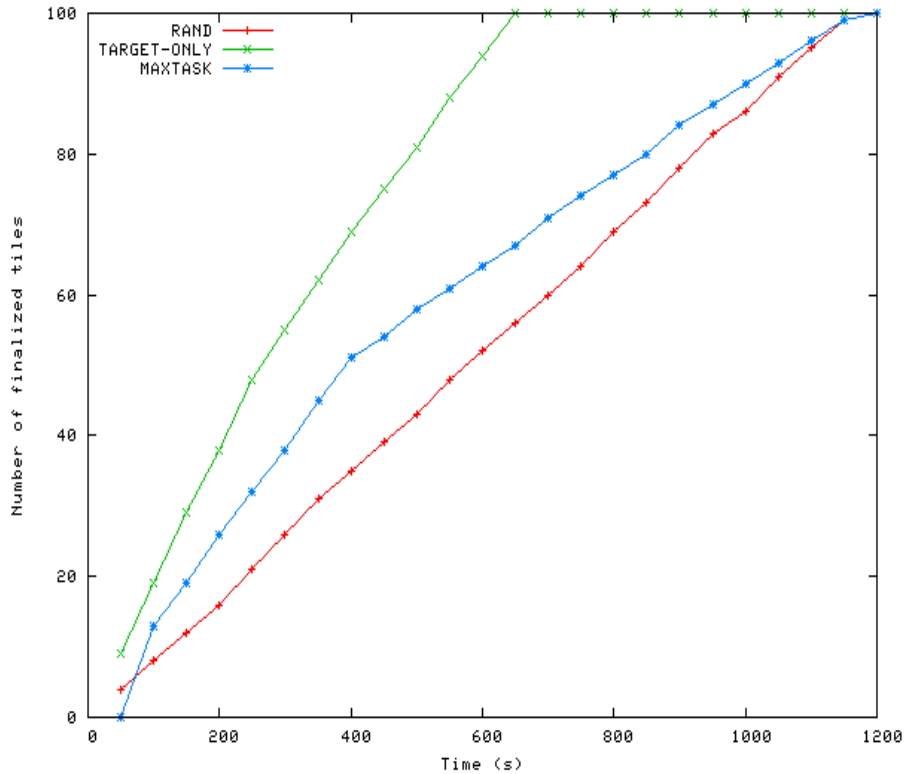
Given a set of image tiles and a user QoS requirement:

- Determine *an* optimal order of execution of the tiles
- Determine the resolution at which each tile must be processed

Simulation studies – Ideal conditions

- Processing time & confidence for tile at each resolution known *a priori*
- Baseline performance (RAND):
 - Pick a tile at random, process tile until it is finalized
- QoS 1: Maximize number of finalized tiles within time t' :
 - Process tiles in **increasing order of their processing times**
- QoS 2: Maximize average confidence for image within time t'
 - Process tiles in decreasing order of **maximum gain in confidence per unit processing time**

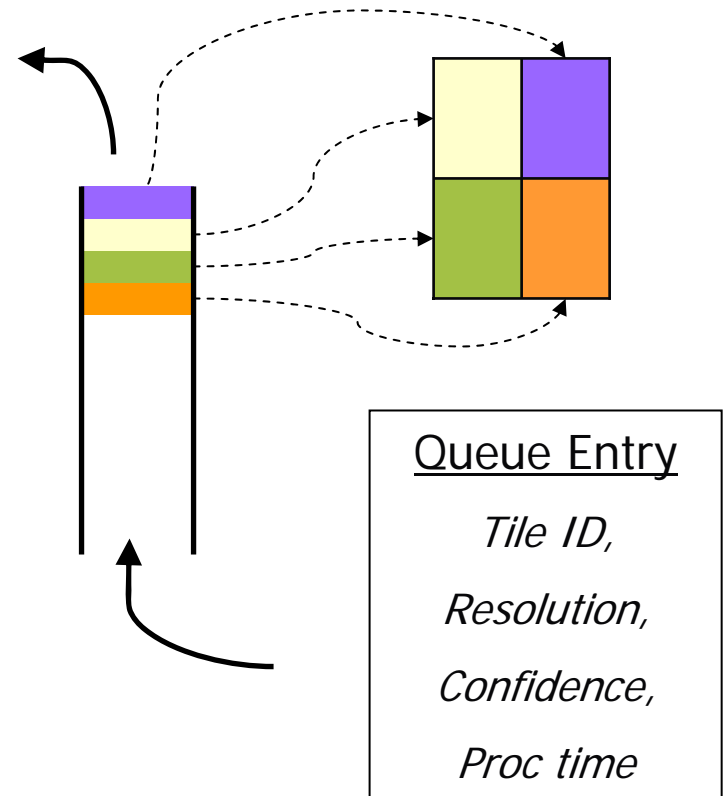
Lessons learnt



- Scope for improvement
- Ordering tile execution can help bridge the gap to a certain extent
- How do we account for non-ideal conditions?

Tile-based heuristic

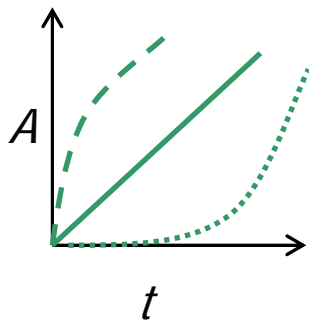
- Initial Static Phase (Only insertions)
 - Process each tile at lowest res
 - Insert entry into Queue
- Dynamic Phase
 - Process entry at top of queue at next higher resolution
 - If tile gets finalized, delete queue entry
 - Insertions and Deletions
 - Queue is dynamically updated
- Use same Priority Queue data structure for different QoS requirements
- Modify insertion operation accordingly



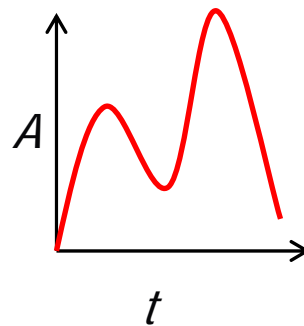
Priority Queue for Tile ordering

Priority Queue Insertion scheme

- QoS: Maximize number of finalized tiles
 - Tile with *confidence* value closer to threshold gets higher priority
 - Better chance of being finalized at *resolution+1*
- QoS: Maximize average accuracy across tiles
 - Tile achieving higher increase in *confidence* per unit of processing time between *resolution-1* and *resolution* gets higher priority
 - Prospect for greater benefit at *resolution+1*



Well-behaved
characteristics

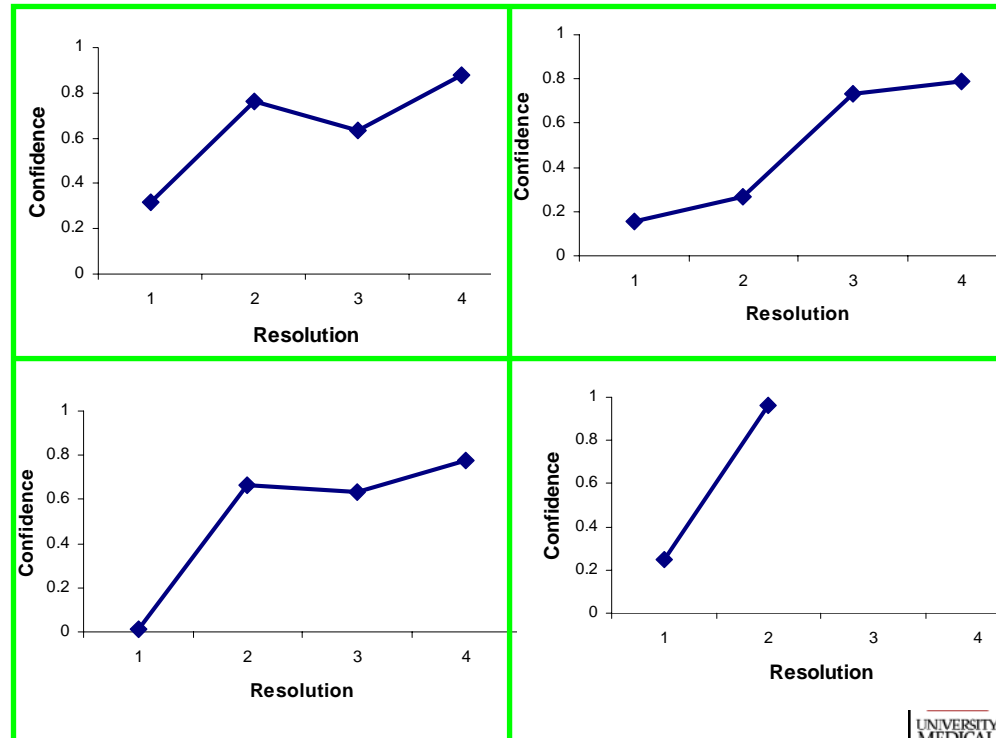
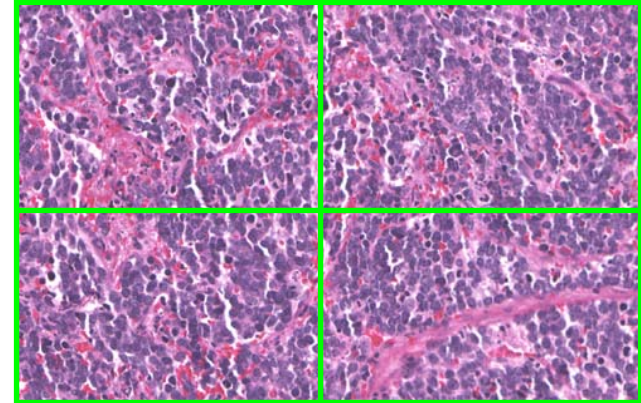


Unpredictable
characteristics

In practice, accuracy
vs. performance
could be non-
monotone in nature

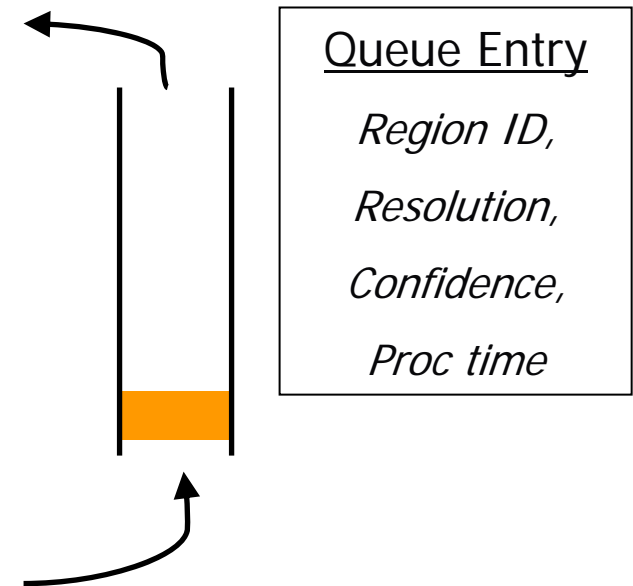
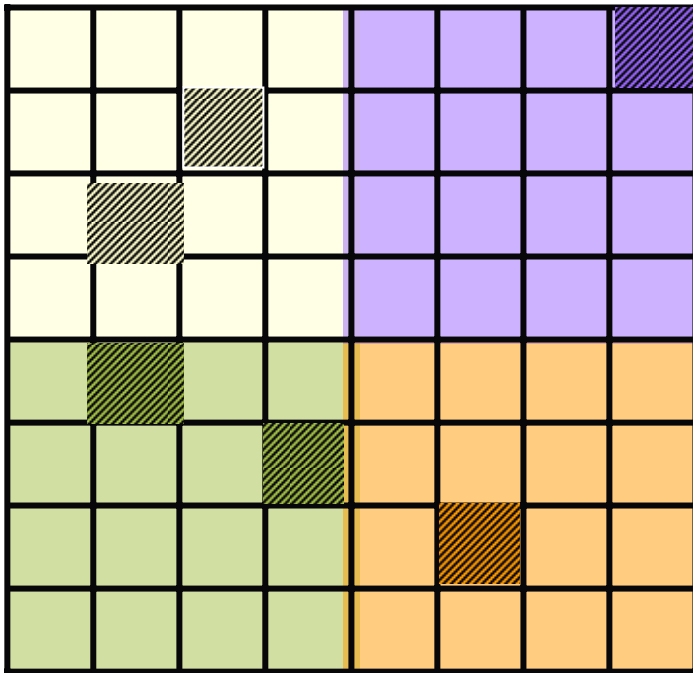
Accuracy vs. performance characteristics

- > 70 % of tiles showed non-monotone characteristics
- Tile-based heuristic is not suited for such cases
 - Impossible to estimate behavior for single tile in isolation
- Spatially close tiles → Similar features → Similar accuracy-performance characteristics
- Group neighboring tiles into *regions*
- Region characterized by its representative tiles.
- Hierarchical region-based heuristic



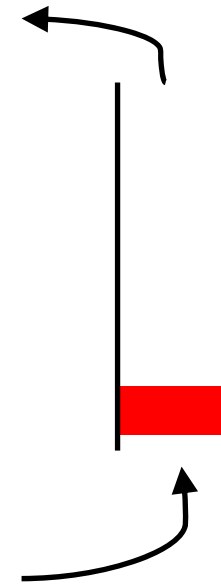
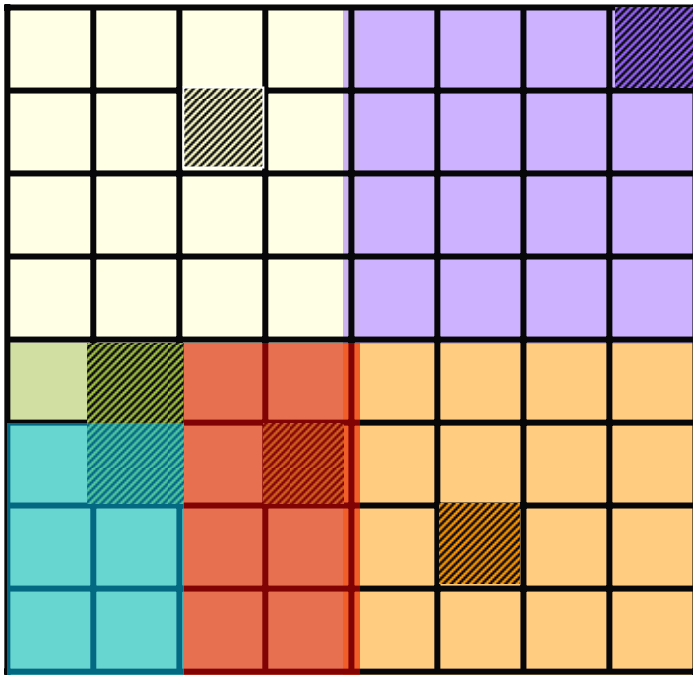
Hierarchical Region-based algorithm

- Uses same priority queue structure. Process a tile completely.
- Queue entry now corresponds to a rectangular *region*.
- Static partitioning approach:
 - Choice of region dimensions: one size will not work for all images
 - Simple, not efficient: does not help converge on favorable tiles

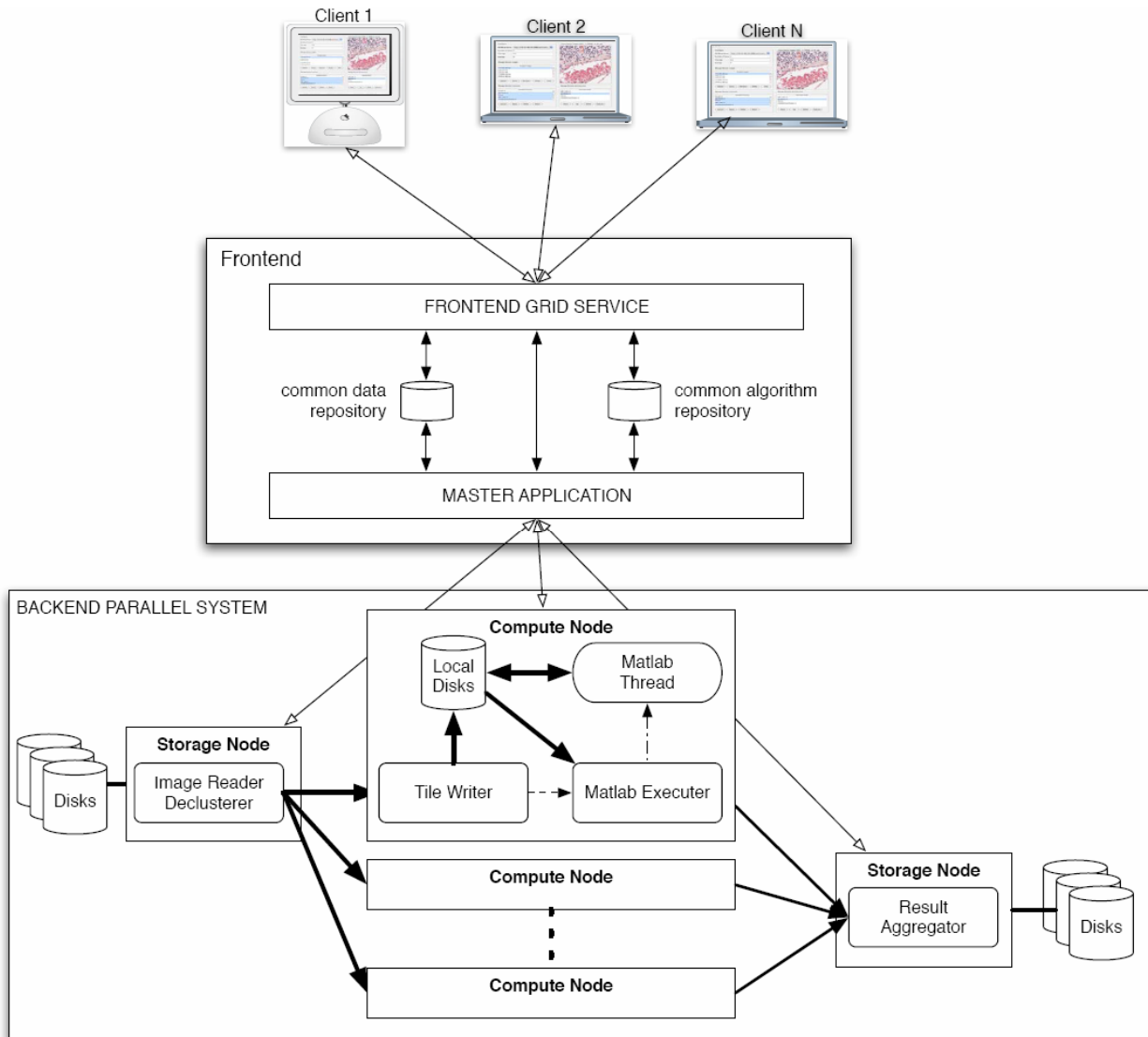


Hierarchical Region-based algorithm

- Uses same priority queue structure. Process a tile completely.
- Queue entry now corresponds to a rectangular *region*.



Architecture

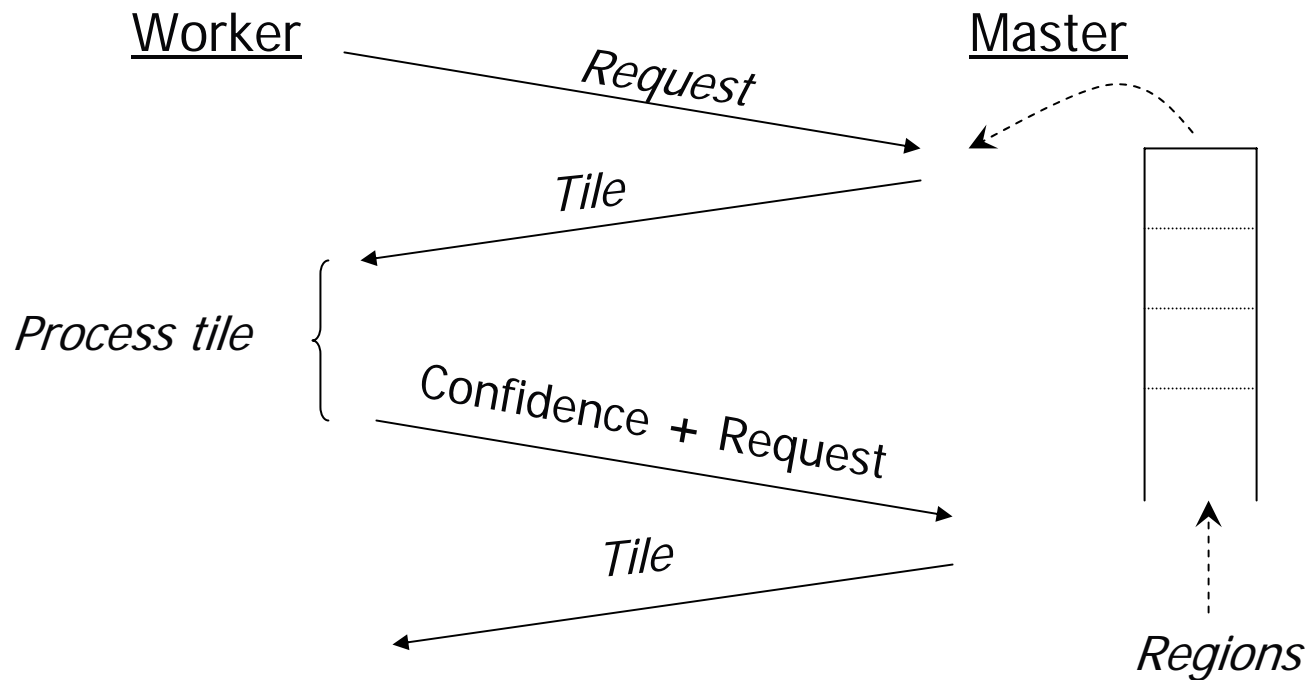


Middleware support:

DataCutter

- Component-stream framework
- Ideal for data analysis workflows
- Combined task- and data-parallelism
- MPI underlying substrate

Parallel Execution



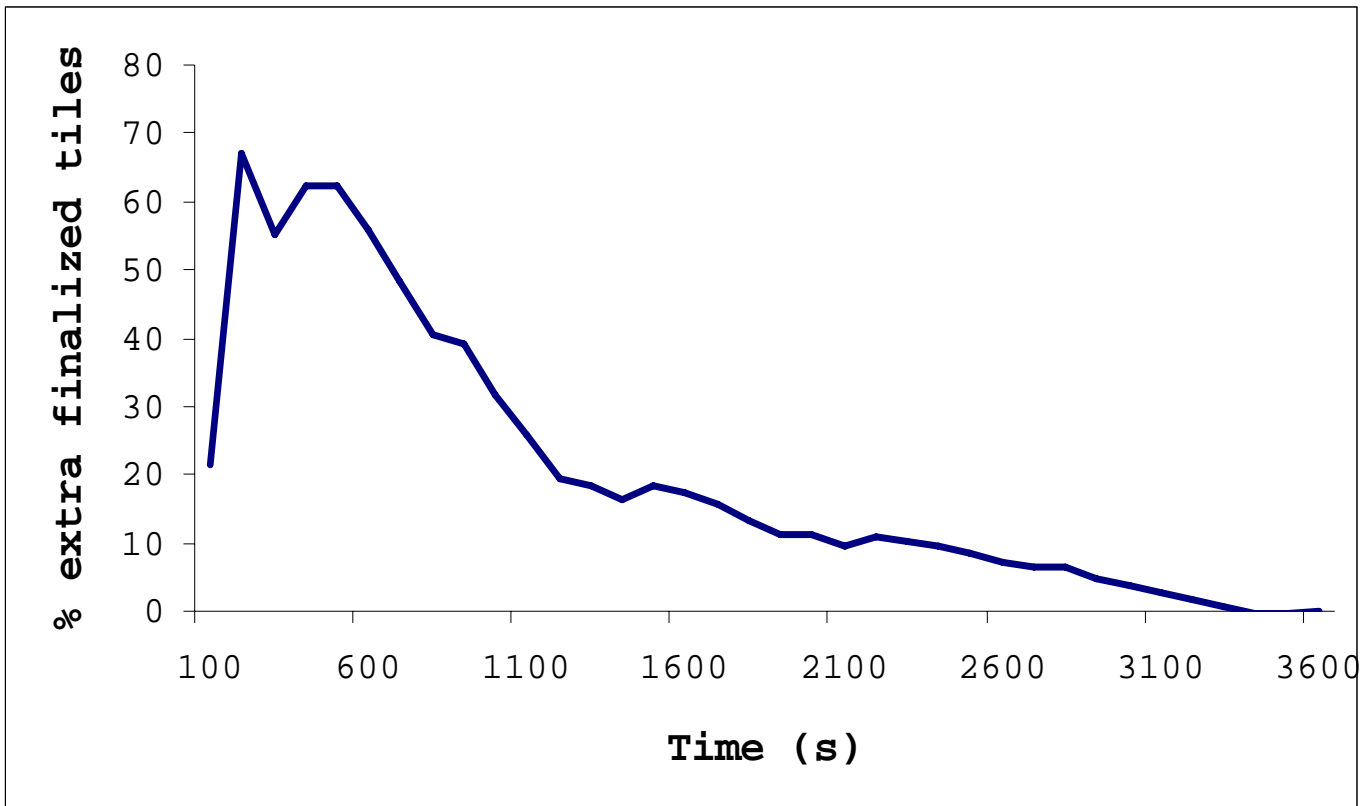
- Static assignment: Leads to load imbalance
- with dynamic redistribution: Higher inter-processor communication
- Demand-driven strategy:
 - Assign 'next' tile/region to worker with earliest request
 - Dynamically adapts to changes to load on workers and network traffic

Experimental test bed

- Cluster of 32 nodes (NSF Research Infrastructure Grant)
- 2.4 GHz AMD Opteron dual-processor nodes
- 8 GB of memory
- 2x250GB SATA disks locally installed on each node (Joined into a 437GB RAID0 volume)
- Interconnected by both an Infiniband and 1Gbps Ethernet network
- Datasets
 - Image 1: 69,840x59,359 pixels (11GB); 4,075 tiles
 - Image 2: 108,640x70,601 pixels (22GB); 16,011 tiles

Quality aspects: Experiment 1

- QoS: Maximize number of finalized tiles within time t'

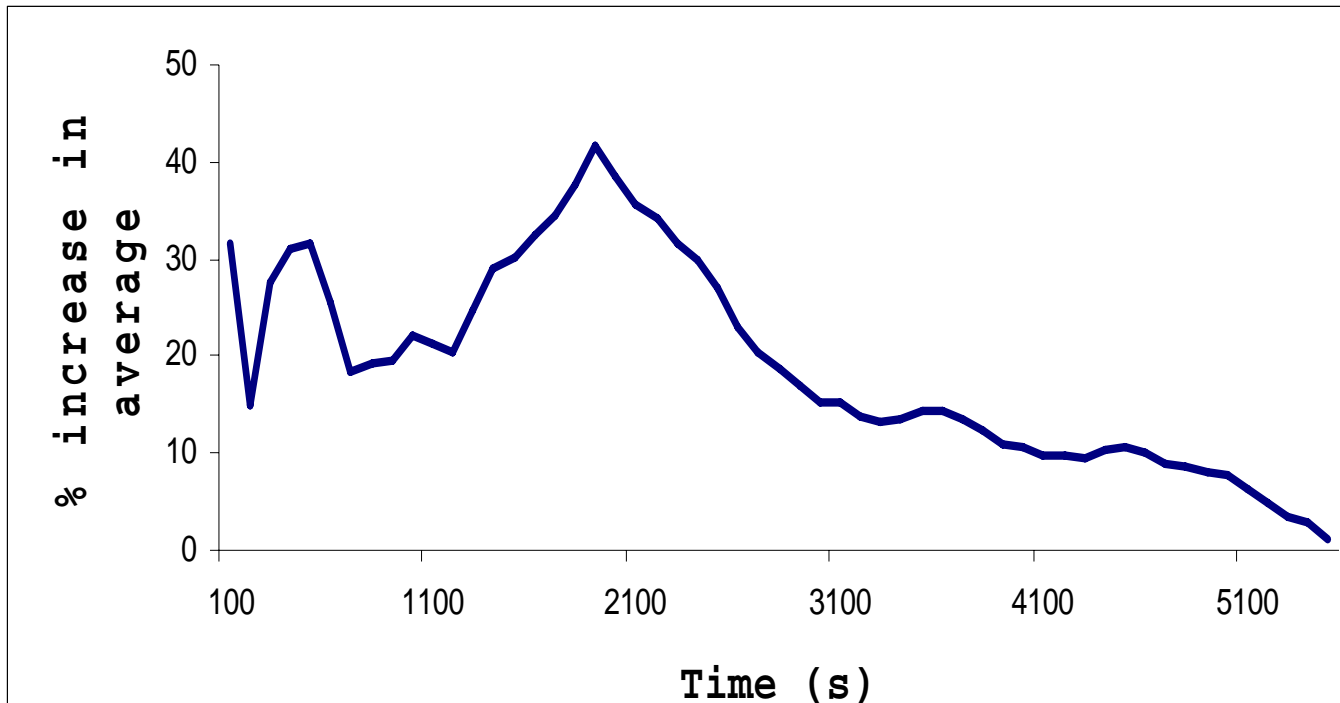


- Image 1
- 16 workers
- Threshold: 0.5

Hierarchical region-based heuristic can finalize 21% more tiles than baseline scheme on an average

Quality aspects: Experiment 2

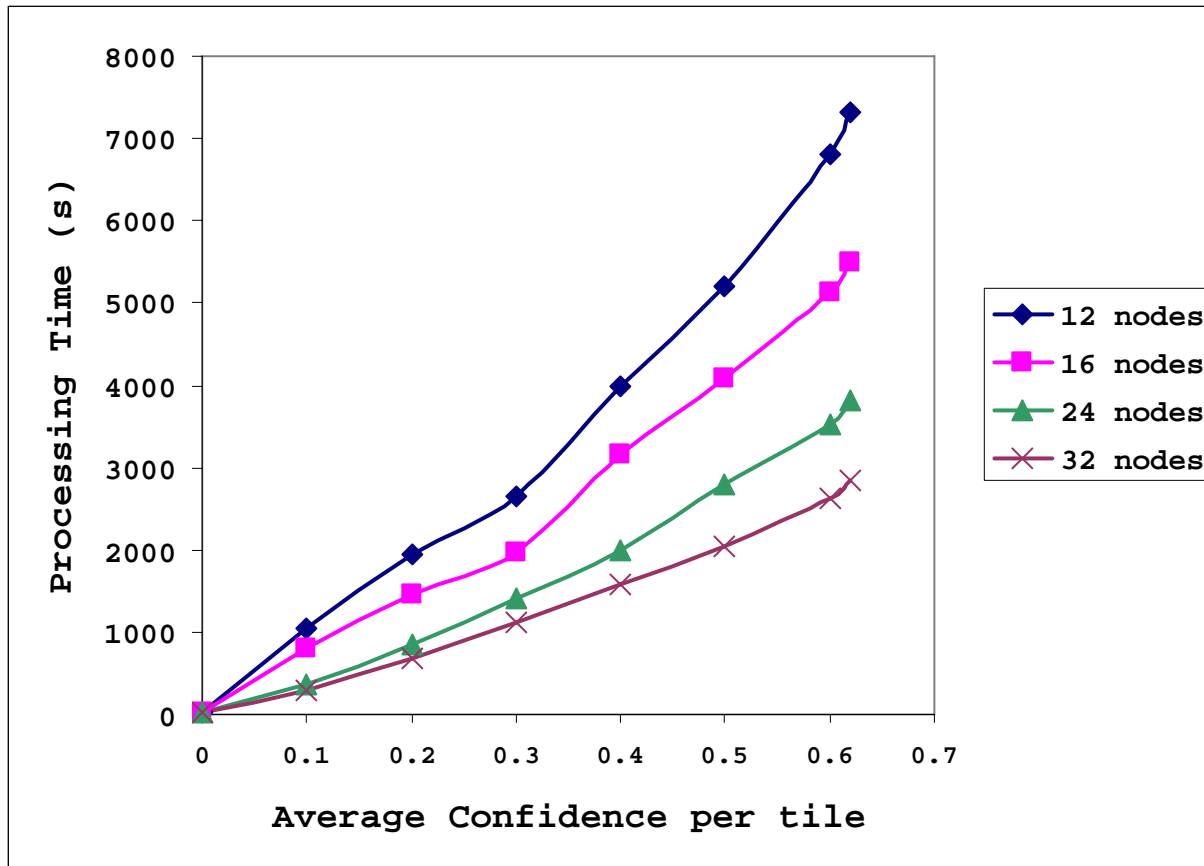
- QoS: Maximize average accuracy across whole image within time t'



- Image 1
- 16 workers
- Threshold: 0.7

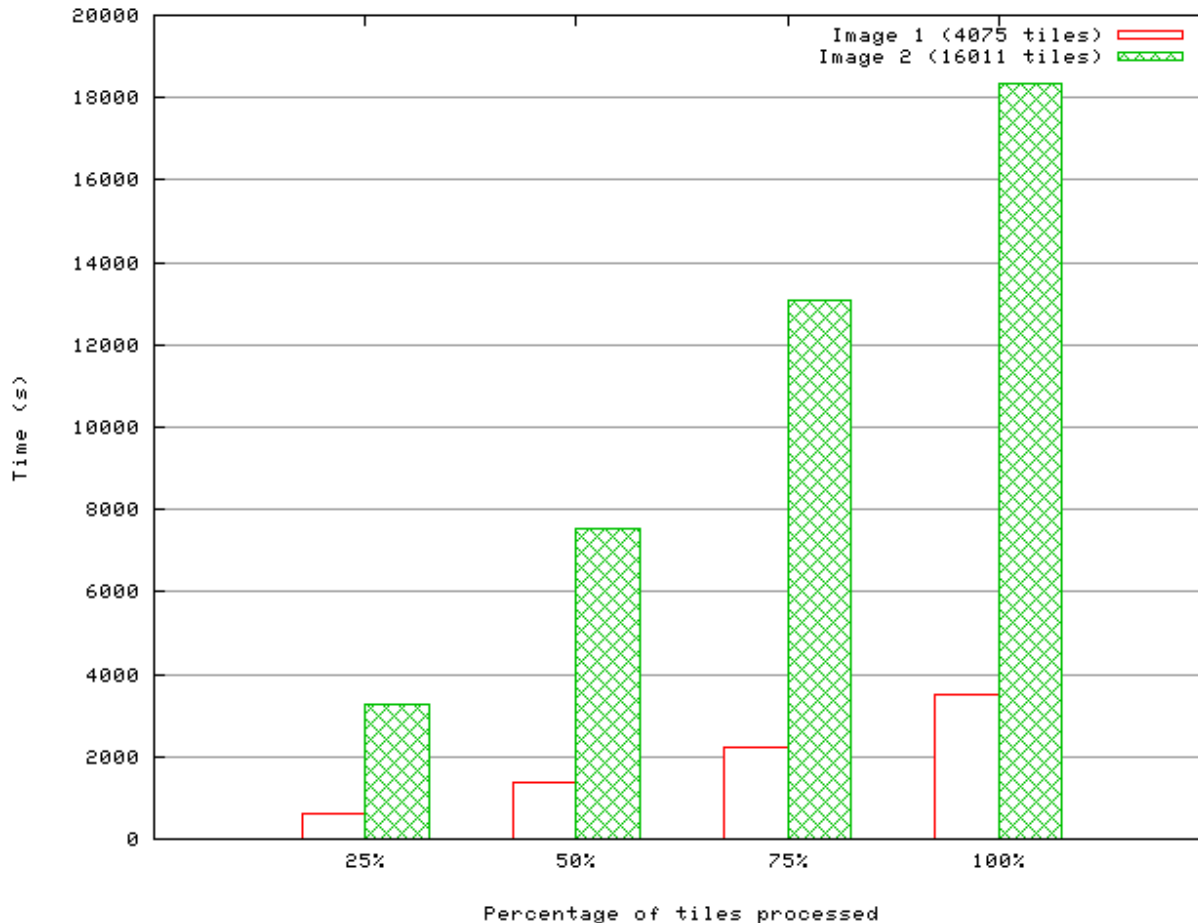
On an average, the confidence for the hierarchical region-based heuristic is 19% greater than that of the baseline scheme

Performance tests: Scaling number of nodes



As number of nodes is doubled, region-based heuristic achieves same average confidence in half the time

Performance tests: Scaling image size



Across different image sizes, linear scaling is observed, but not strictly linear.

Related Work

- Large-scale Image Analysis
 - Virtual Microscopy: BrainMaps.org (UC Davis)
- Support for distributed adaptive applications
 - Language and compiler-based approaches (Adve et al, Agrawal et al)
 - Prediction of execution time
 - Assume data invariance
 - Middleware-based approaches
 - ActiveHarmony, ERDoS, EPIQ, CQoS
 - Focus on time-related requirements
 - Adaptation to changes in resources and execution environment
 - Fine-grained adaptation (at a single data-component level)
- Our framework can be complemented with benefits from above approaches

Conclusions

- Framework for distributed adaptive processing of image tiles
- Support for different user QoS requirements
- Tile-ordering heuristics help bridge gap between basic and optimal execution strategies
- Scales well with compute nodes and image data size.

- Ongoing work:
 - Express application parameters and QoS requirements to the system
 - Improve heuristics depending on application characteristics
 - Realistic user requirements

Performance vs. Accuracy Trade-offs for Large-scale Image Analysis Applications

Thank you!

<http://www.bmi.osu.edu>

vijayskumar@bmi.osu.edu