NORTHEASTERN
STATE UNIVERSITY

Cloud-based Computing

Neal N. Xiong, Ernst Bekkering, Steven Rice
Northeastern State University, USA

Outline

- A Basic Idea: cloud, IoT, SDN, ...
- Rainfall Estimation Systems
- Data Fusion in Mobile Wireless Sensor Networks
- NSU Computer Science Beowulf Computing Cluster

NORTHEASTERN STATE UNIVERSITY

Evolution of HPC, Clusters, Distributed P2P/Grid/**Cloud** Computing, and Web Services

Disparate systems sharing

Distributed Computing

High Perf. Computing

Similar Proc.

Results in High Performance

Mainly for file sharing

Resource Sharing

P2P

Clusters

Close to each other

Geographically Sparse
No existing framework

Geographically Sparse
Within a Framework

Disparate Systems
Resource Sharing,
Geographically Sparse
Within a Framework

Grid/Cloud

Web Services

Various Application Interaction

NORTHEASTERN STATE UNIVERSITY

Internet of Things – Introduction

*The Internet of Things (IoT) is the **network of physical objects** that contains embedded technology to **communicate** and **sense or interact** with the objects' **internal** state or the **external** environment.**

*Gartner,July 2014

NORTHEASTERN STATE UNIVERSITY

What's hindering IoT?

Fragmentation

- Multiplicity of standards
- No cross-industrial standard / reference design
- Interoperability issues at different levels of software and hardware

Software, Services and Algorithms

- Lack of open IP-Based Connectivity
- Delay in IPv6 deployment
- Low Power "Self-healing sensor Network"
- Efficient protocol and topology to form Heterogeneous Mesh Networks
- Creating flexible and scalable systems

Security & Privacy

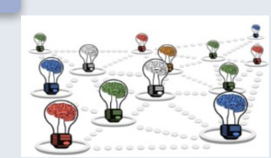
- Secure and authorized access
- Privacy of User Data
- Models for decentralized authentication
- Data integrity at consumer devices

Device & Data Management

- Seamless plug-n-play of sensor nodes
- Managing millions of connected smart objects real time event handling
- Handling the huge data generated by IoT devices

Collective Intelligence

- Individual nodes and sensors may not be "Smart" enough
- Need to leverage big data and cluster of sensors/device to bring in collective intelligence



NORTHEASTERN STATE UNIVERSITY

Software-Defined Networking

An Effective Rainfall Estimation System

NORTHEASTERN STATE UNIVERSITY

Background

Unfortunately, quantitative precipitation estimation is challenging around the world!



- emergency rescue and disaster relief
- large scale construction
- high speed rail operation support



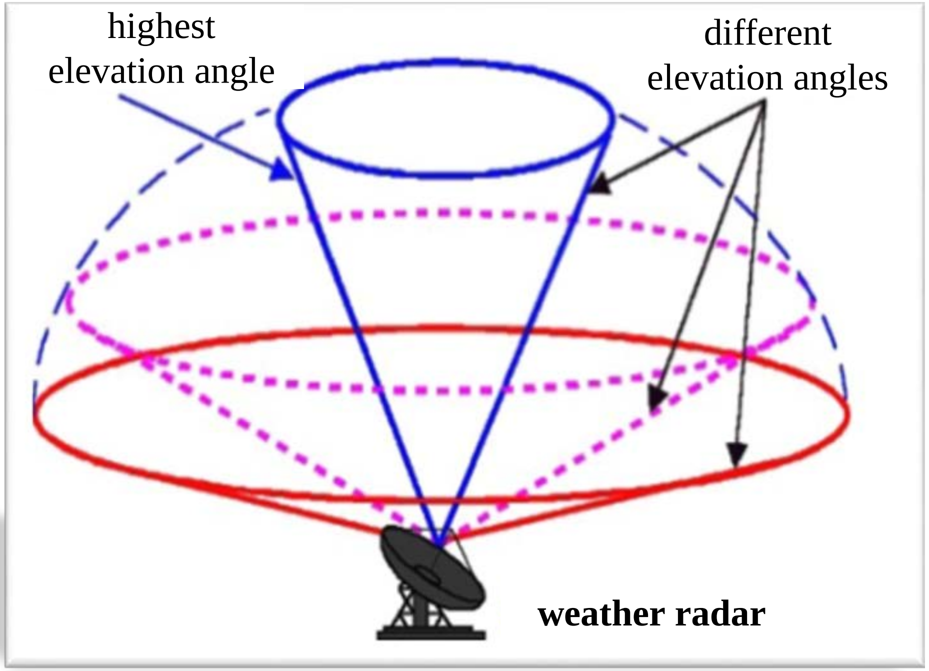
- spa… lau…
- … activity

accurate precipitation estimation is required

Era of weather radar big data

Volume	larger than 1PB per year.
Velocity	update 6 million observed data for each radar in 6 min.
Variety	radar reflectivity, radial velocity, spectral width, etc.
Value	rainfall estimation, hail recognition, tornado, etc.
Veracity	suffer from various factors, high uncertainty.

radar big data



spatial and temporal resolution
1km×1km×6min

Challenges

1. **Affected by many factors**

   （Sivakumar B et al.,2015）

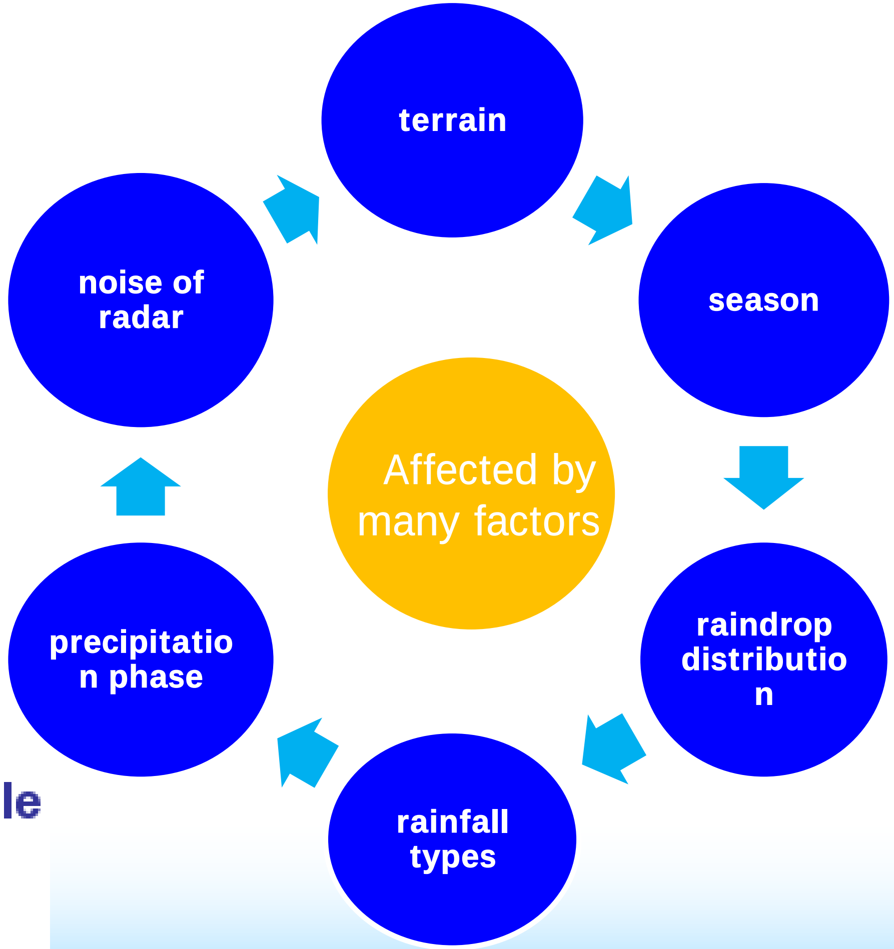
2. **Challenging problem**

   - **Hydrology** (Li, P.W.,2004)
   - **Data mining**（Yang, Y et al.,2007）
   - **Environment-related machine learning**

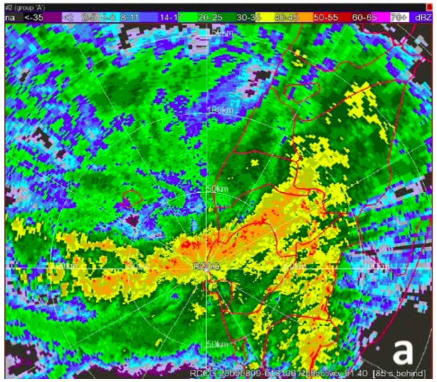
     (Hong,W,2008)
   - **Statistic forecast** (Pucheta, J., 2009)

3. **Data-deriving modeling is a possible solution**

   (C.L. Wu and K.W. Chau ,2013)

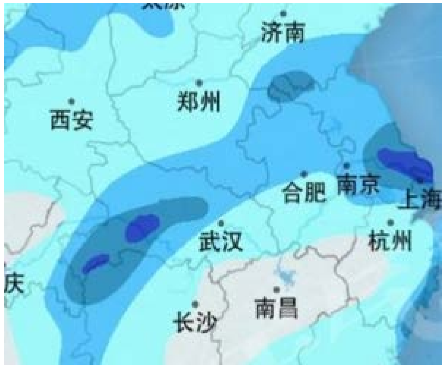


Conventional radar rainfall estimation



estimate

$$Z = aR^b$$



radar data

rainfall data

convective

$Z = 320R^{1.44}$

stratiform

$Z = 200R^{1.6}$

Difficult to meet ideal condition

- Raindrop fall with equivalent speed
- low speed of wind
- uniform distribution of raindrop

Attempt in machine learning perspective

$$\log(Z) = b\log(R) + \log(a)$$

regression

- **Xiao *et al*. TGRS 1997**

  output R
  by NN

Drawback: Existing ML methods directly treat this problem as regression in statistics.

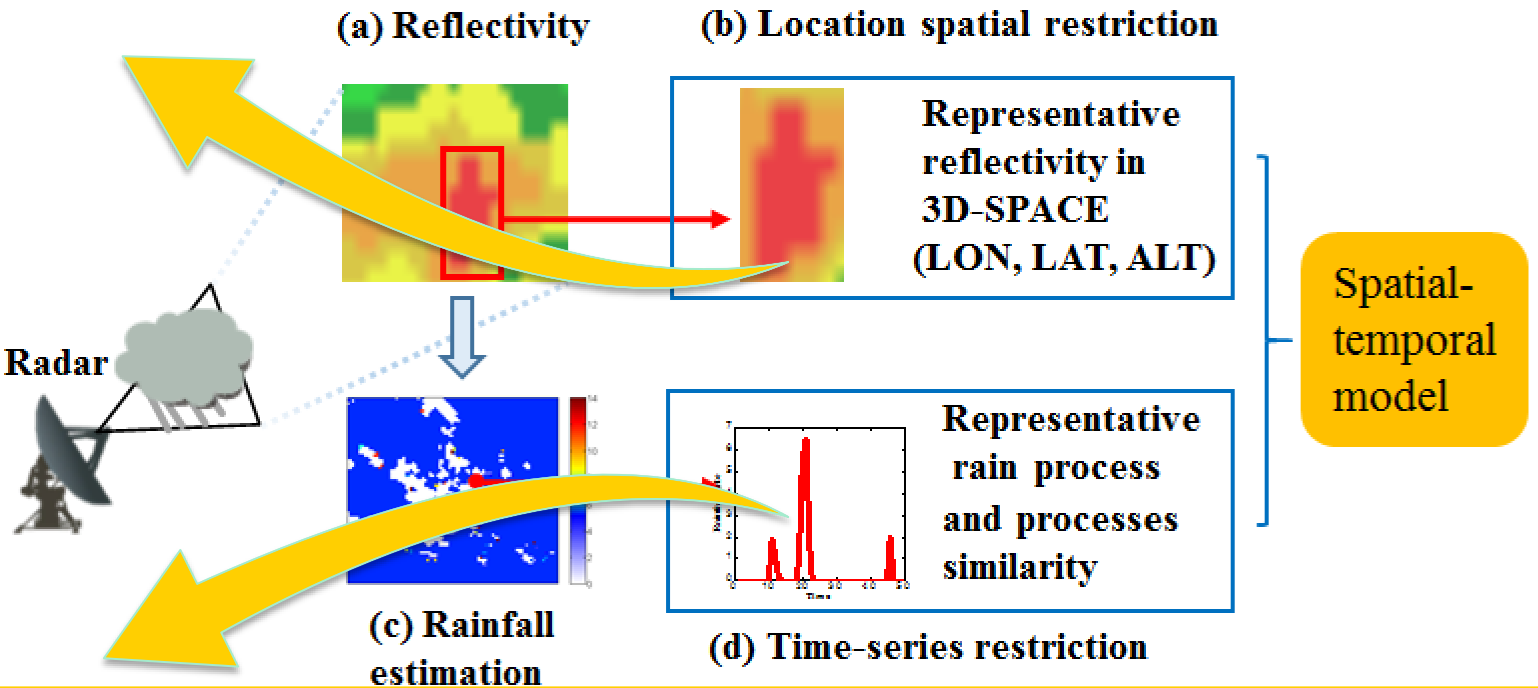
- **Huang *et al*. IIDS 2015**

  input 1,2,3,4 km Z & radial velocity

- **Kusiak *et al*. TGRS 2013**

RANLIST model

spatial structure $\min_{\Theta} f(\Theta, x', y) = \sum \left(\sum_{v \in S_j} \left(\left[y(\Theta) - \bar{y}_j(\Theta) \right]^2 \right) - \sum_{i \in L,R} \sum_{v \in S_j} \left(\left[y(\Theta) - \bar{y}_j(\Theta) \right]^2 \right) \right)$

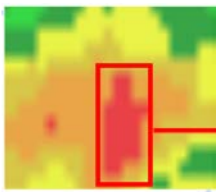


temporal structure $\min_{\omega,\lambda} g(\omega, \lambda) = \sum_{k=1}^{p} \left[\sum_{j=1}^{q(k)} \log\left(Z_k(x)\right) - \sum_{j=1}^{q(k)} \sum_{t=1}^{T} \omega_{kt} P_{kj}\left(y_{t-1}, y_t, x_t\right) + \lambda \left\| \omega_k \right\|_1 \right]$

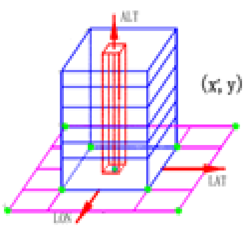
Qiuming Kuang, Xuebing Yang, Wensheng Zhang, Guoping Zhang. Spatiotemporal Modeling and Implementation for Radar-Based Rainfall Estimation[J]. IEEE Geoscience and Remote Sensing Letters, 13(11):1601-1605, 2016. （SCI，IF: 2.228）

RANLIST model

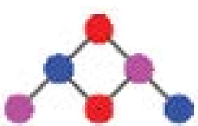
$$\min_h \boxed{L_S(h(x,y)) + \lambda_1 L_D(h(x,y)) + \lambda_2 L_{SP}(x',y')} + \boxed{\lambda_3 L_t(x, y_{t-1}, y_t)}$$



Reflectivity



Spatial extention



RF

Spatial

$$\min_{\Theta} f(\Theta, x', y) = \sum \left(\sum_{v \in S_j} \left([y(\Theta) - \bar{y}_j(\Theta)]^2 \right) - \sum_{i \in L,R} \sum_{v \in S_j^i} \left([y(\Theta) - \bar{y}_j(\Theta)]^2 \right) \right)$$

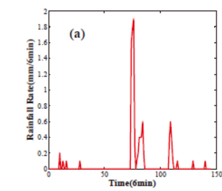
Spatial submodel RANMP

$$\min_{\omega,\lambda} g(\omega,\lambda) = \sum_{k=1}^{p} \left[\sum_{j=1}^{q(k)} \log\left(Z_k(x)\right) - \sum_{j=1}^{q(k)} \sum_{t=1}^{T} \omega_{kt} P_{kj}(y_{t-1}, y_t, x_t) + \lambda \|\omega_k\|_1 \right]$$

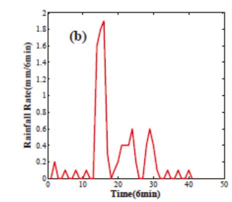
Time-series submodel LITS

$$\begin{aligned} \min_{\Phi} \quad & K_{s1} K_{r1} f(\Theta) + K_{s2} K_{r2} g(\omega,\lambda) \\ \text{s.t.} \quad & K_{r1} + K_{r2} = 1 \end{aligned}$$

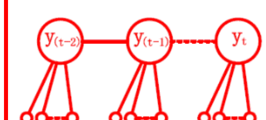
Spatiotemporal model RANLIST



Time-series rainfall



Rain processes



LCCRF

temporal

NORTHEASTERN STATE

RANLIST——Three-stage optimization

$$\min_{\Phi} \; K_{s1}K_{r1}\, f(\Theta) + K_{s2}K_{r2}\, g(\omega, \lambda)$$

[1] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

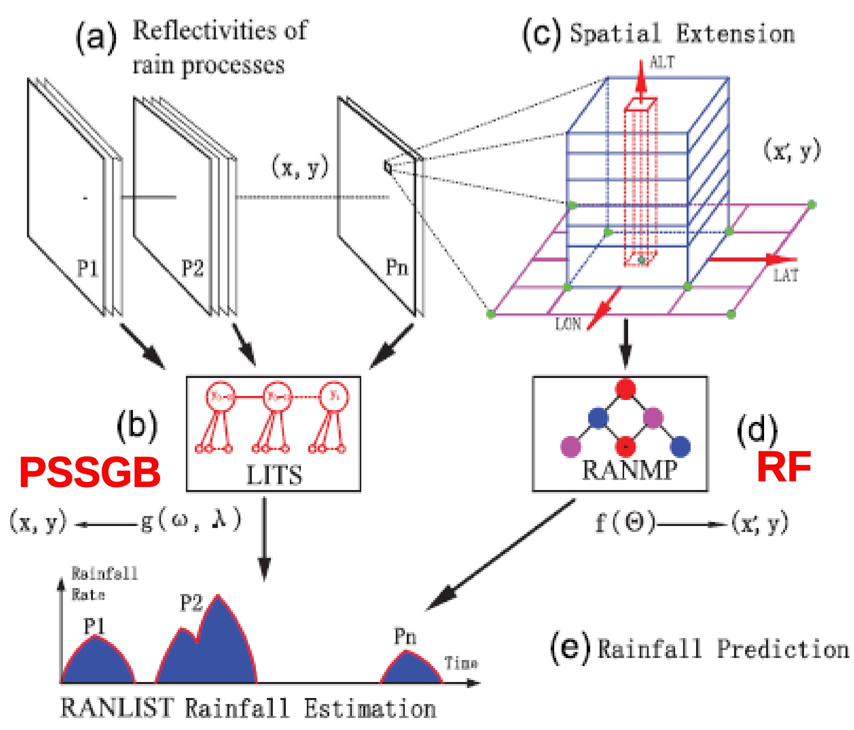
[2] M. Schmidt, "Graphical model structure learning with l1-regularization," Ph.D. dissertation, Univ. British Columbia, Vancouver, BC, Canada, 2010.

Three-stage

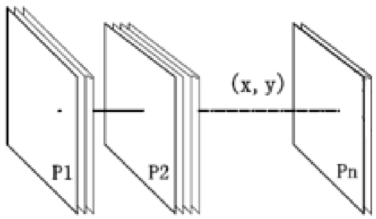
First-stage : Random Forest method for optimizing $f(\theta)$

Second-stage : PSSGB method for optimizing $g(\omega, \lambda)$

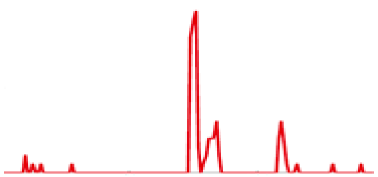
Three-stage: : $K_{s1}, K_{s2}, K_{r1}, K_{r2}$ are determined by validation data set



Training RANLIST model



Reflectivity factors



Time-series rainfall

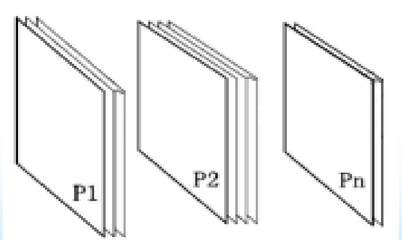
Quality Control → Multiple CAPPI Reflectivity → Spatial Interpolation

Temporal Interpolation → Data Matching

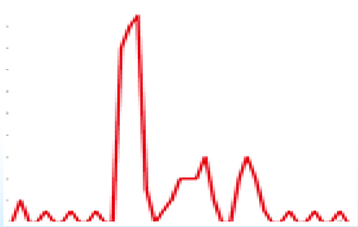
Data preprocessing

Accroding to rainfall

Obtain rain processes



Reflectivity factors for each rain process



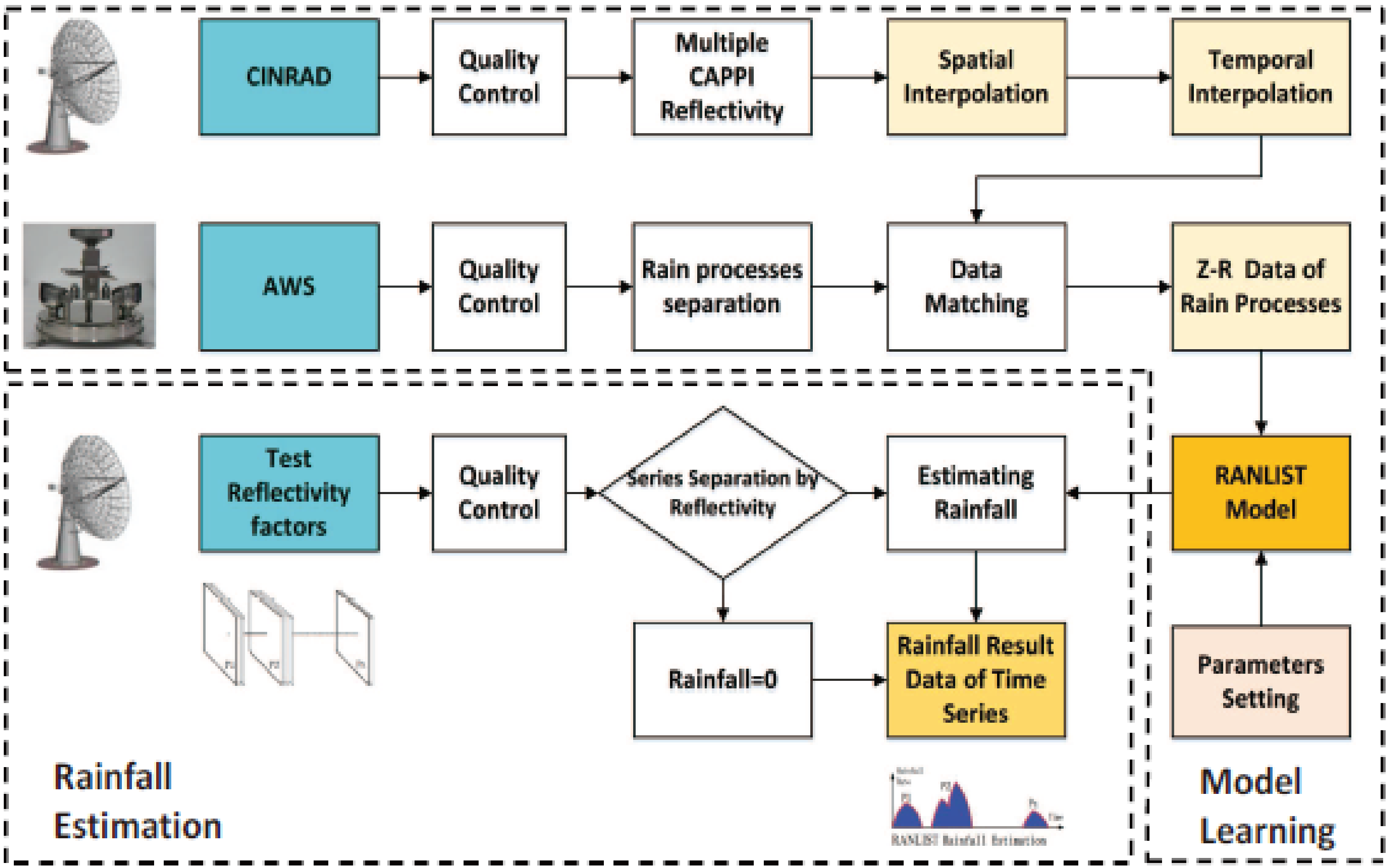
Rainfall for each rain process

Model training by rain processes

RANLIST

NORTHEASTERN STATE UNIVERSITY

Workflow of RANLIST model

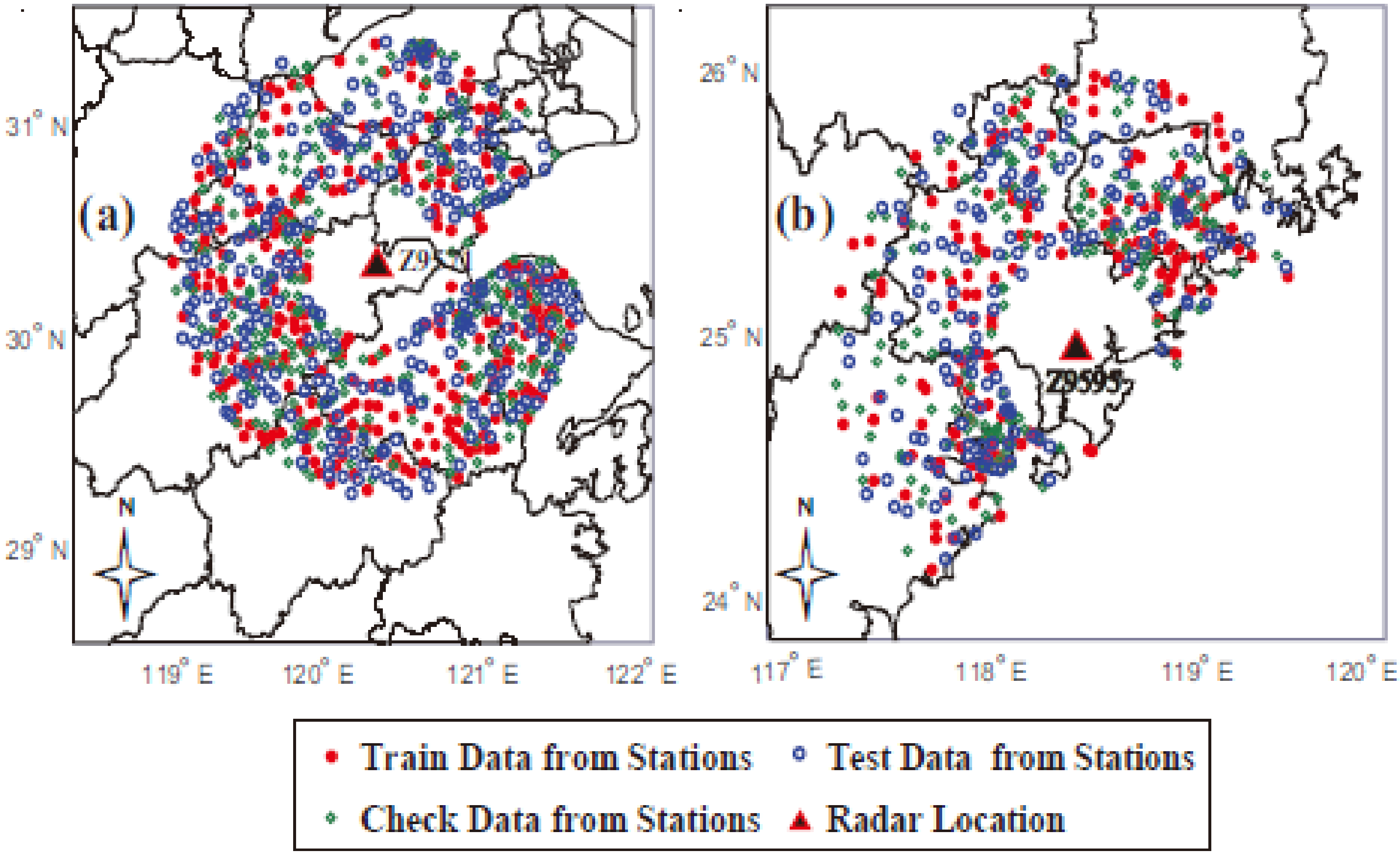


Data sets for experiments

DATA SETS DESCRIPTION

Data Sets	region	days	total number of gauge-radar pairs
Data Set 1	Hangzhou	21th,25th,26th June and 12th,13th,15th July,2014	1,124,640
Data Set 2	Quanzhou	16th June and 23th,24th July,2014	336,960
Data Set 3	Quanzhou	16th June and 23th,24th July,2014	58,950

Experimental areas



Result

Comparison between radar rainfall estimation and raingauge measurement for Hangzhou radar coverage

	$RMSE_{mm}$	MAE_{mm}	CC
Z-R	3.05	1.64	0.659
SVR	2.69	1.51	0.717
RF	2.80	1.41	0.718
RANMP	2.72	1.36	0.737
LIST	2.51	1.14	0.763
RANLIST	**2.15**	**1.05**	**0.829**

1. Compared to conventional Z-R relationship:
Improvements of **30% in RMSE**, **36% in MAE** and **26% in CC** are obtained

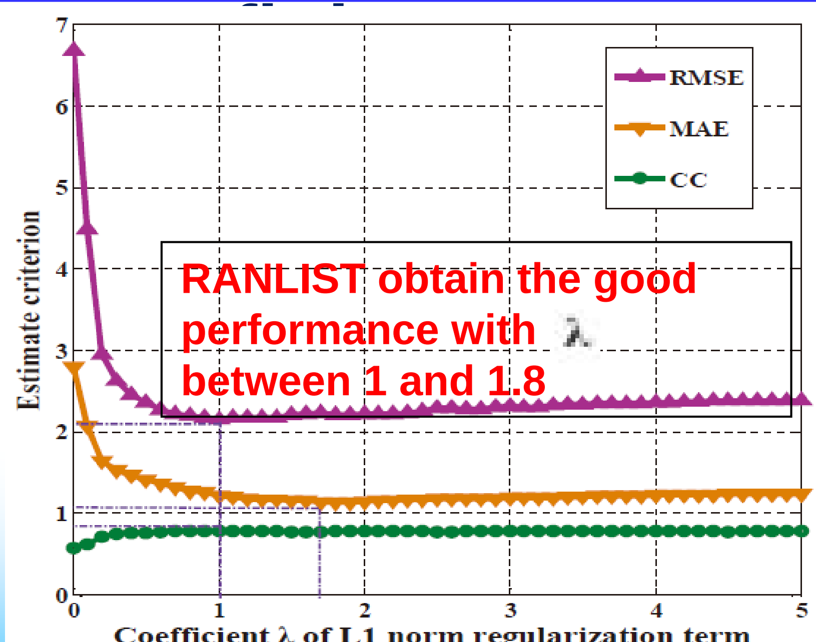
2. Compared to random forest:
Improvements of **23% in RMSE**, **26% in MAE** and **15% in CC** are obtained

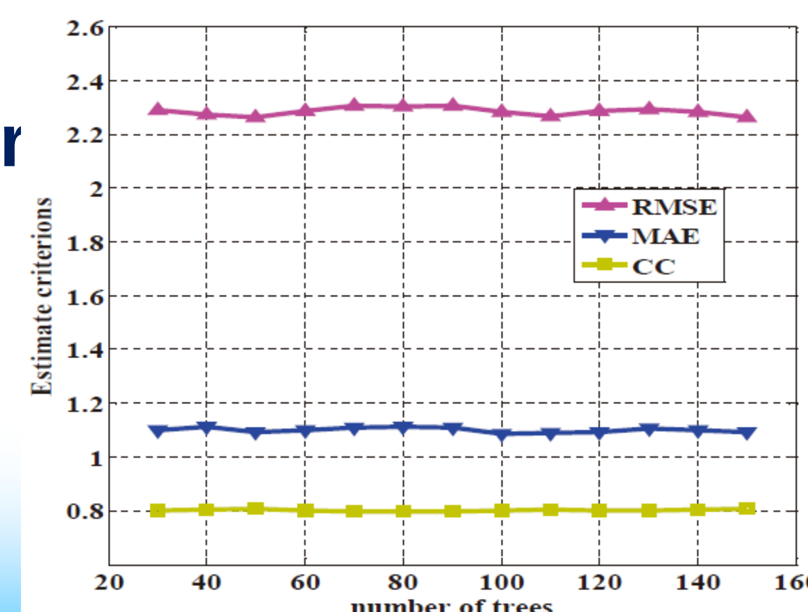
Evaluation

- **Several most similar rain process affect rainfall estimation seriously.**
- **Very good fitting rain processes will lead to overfitting.**
- **Number of trees is larger than 50, there is no obvious differences.**

Evaluation

- Several most similar rain process affect rainfall estimation seriously.
- Very good fitting rain processes will lead to





Evaluation

- **RANLIST model achieves some advantages in both noisy environments.**
- **The performances are affected because a large number of rainfall processes are destroyed by abnormal or missing data.**

Estimation effectiveness of radar based rainfall estimation methods at the region of Hangzhou

	RMSE	MAE	CC
Z-R	2.32	1.04	0.859
SVR	1.80	0.84	0.888
RF	1.76	0.85	0.890
RANLIST	**1.57**	**0.72**	**0.913**

Estimation effectiveness of radar based rainfall estimation methods at the region of Quanzhou

	RMSE	MAE	CC
Z-R	4.70	2.54	0.644
SVR	3.75	2.21	0.715
RF	3.74	2.22	0.718
RANLIST	**3.48**	**2.10**	**0.756**

Evaluation

- IDW(Inverse distance weight) and kriging are two typical interpolation methods.
- LITS submodel can achieve better performance under interpolating method of kriging.
- RANLIST may achieve even good performance using Kriging . However, some more works must be done.

ESTIMATION EFFECTIVENESS OF DIFFERENT SPATIAL INTERPOLATION METHODS

	IDW			Kriging		
	RMSE	MAE	CC	RMSE	MAE	CC
RANMP	2.92	1.97	0.622	3.00	2.11	0.643
LITS	3.48	1.83	0.640	3.14	1.53	0.671
RANLIST	**2.54**	**1.39**	**0.712**	**2.49**	**1.45**	**0.702**

NORTHEASTERN STATE UNIVERSITY

Evaluation

- RANLIST obtain good performance when the error of determining no rain is not considered
- Rain/no rain classification is especially important for 1-min rainfall estimation

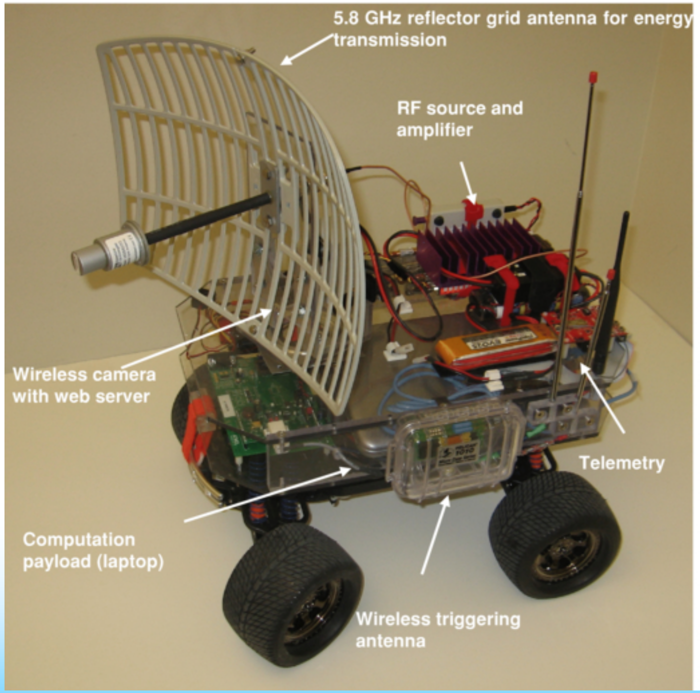
ESTIMATION EFFECTIVENESS OF DIFFERENT TEMPORAL RESOLUTION BY DETERMINING NO RAIN WITH MAX REFLECTIVITY LESS THAN 20DBZ

	6-min			1-min		
	RMSE	MAE	CC	RMSE	MAE	CC
RANMP	2.81	1.82	0.708	6.11	5.11	0.471
LITS	3.39	1.64	0.637	5.88	4.29	0.344
RANLIST	**2.68**	**1.40**	**0.736**	**5.74**	**4.71**	**0.467**

ESTIMATION EFFECTIVENESS OF 6-MIN AND 1-MIN WITH NOT CONSIDERING THE ERROR OF NO RAIN

	6-min			1-min		
	RMSE	MAE	CC	RMSE	MAE	CC
RANMP	**2.78**	1.43	0.746	2.27	1.04	0.832
LITS	3.13	1.53	0.719	2.64	1.00	0.828
RANLIST	2.83	**1.28**	**0.766**	**2.25**	**0.99**	**0.840**

Data Fusion in Mobile Wireless Sensor Networks



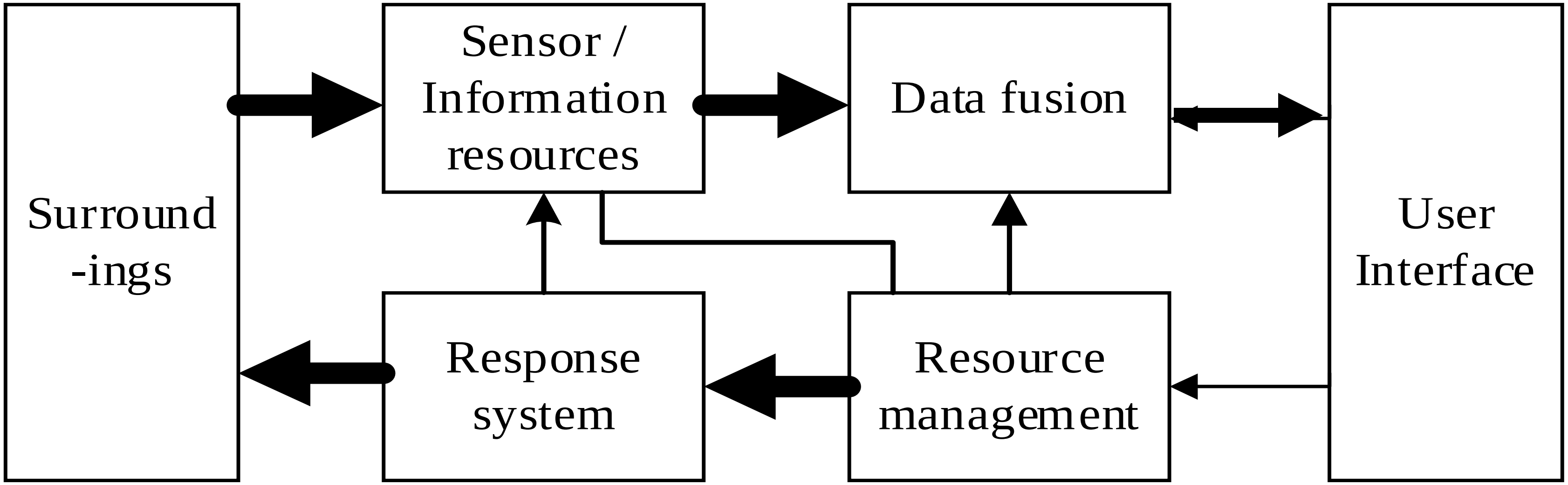
Background

- Mobile sensors
  - Work in different environments
  - Collect data to be fused
  - Have limited energy and storage
  - May work with different protocols
- When events are detected
  - Paths are dynamically selected
  - Data is processed in stages

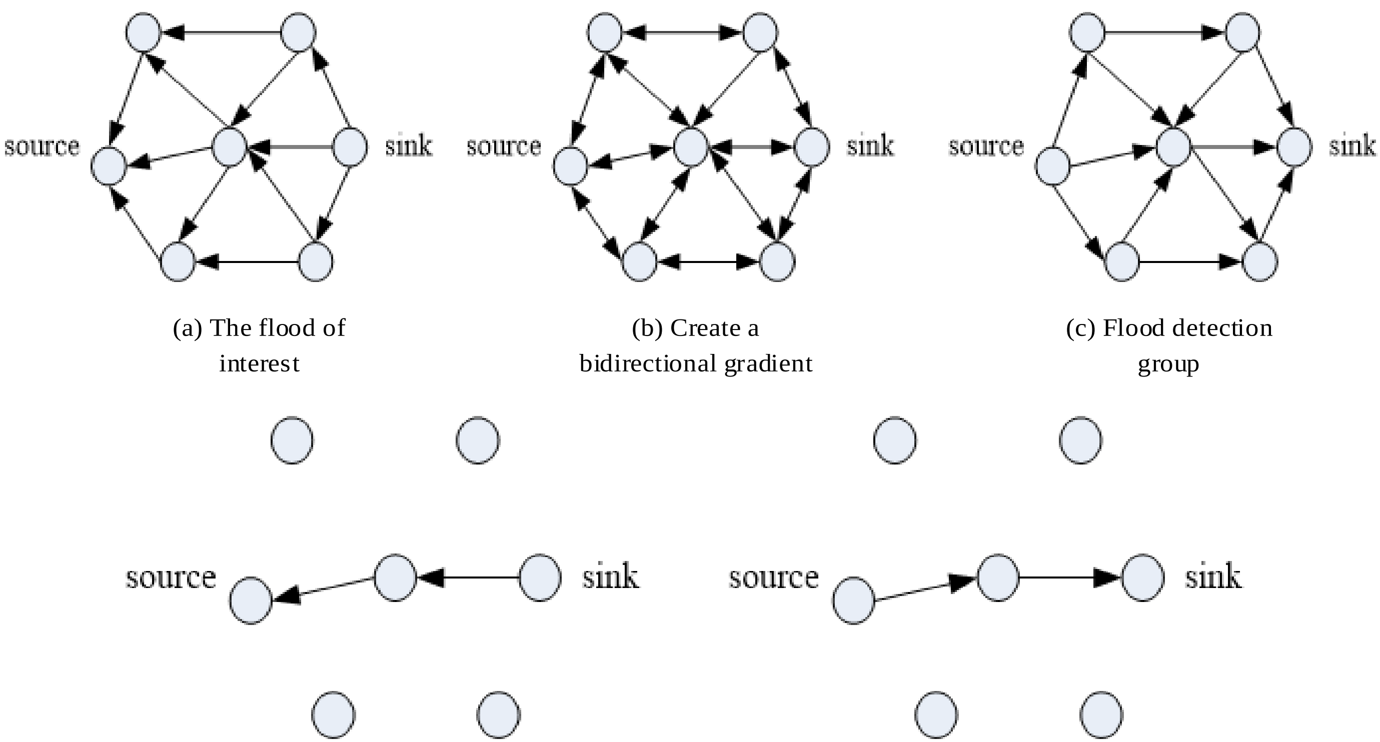
Application area examples

- Monitoring of industrial processes
  - Oil production
  - Chemical industry
- In presence of danger
  - Areas with nuclear radiation
  - Areas with chemical pollution
- Many other areas where data has to be collected autonomously

Data fusion architecture



Routing Protocols : query-based



(a) The flood of interest

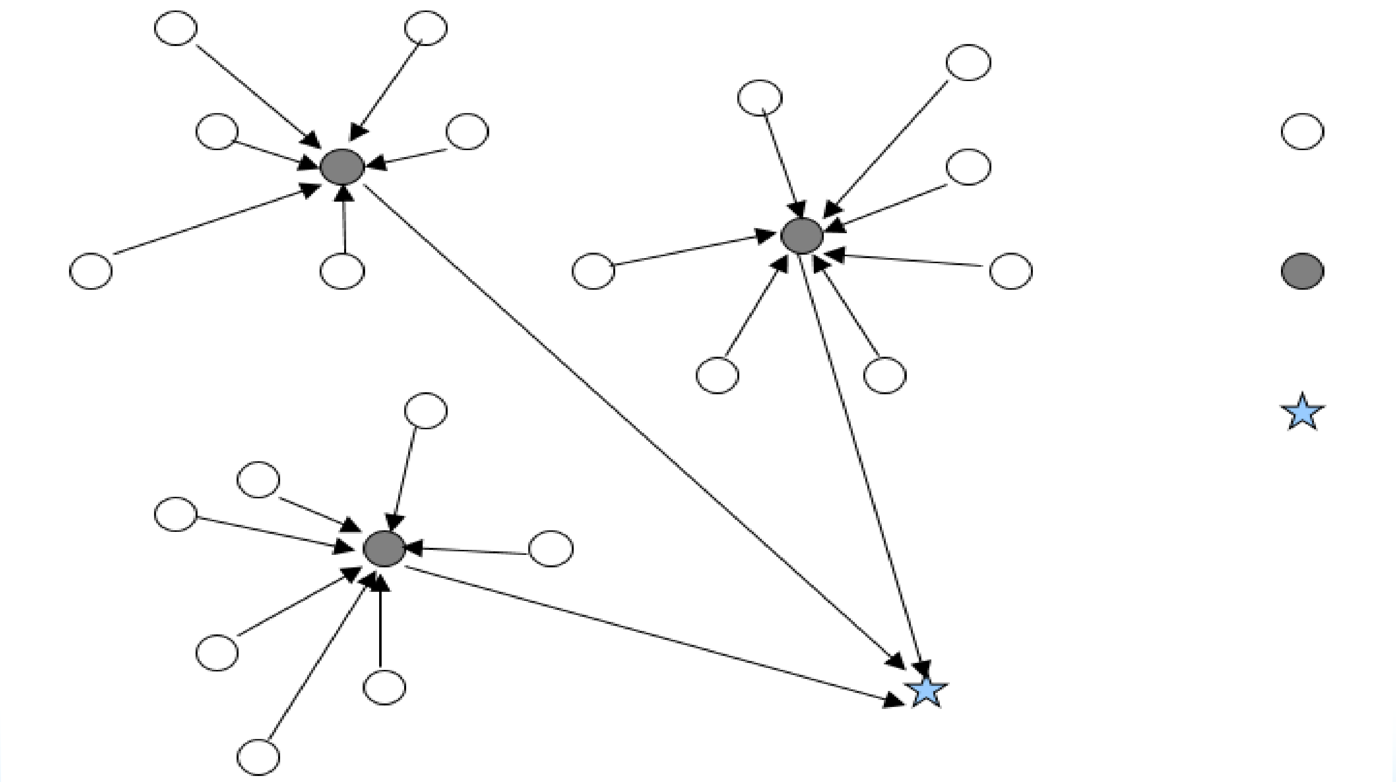
(b) Create a bidirectional gradient

(c) Flood detection group

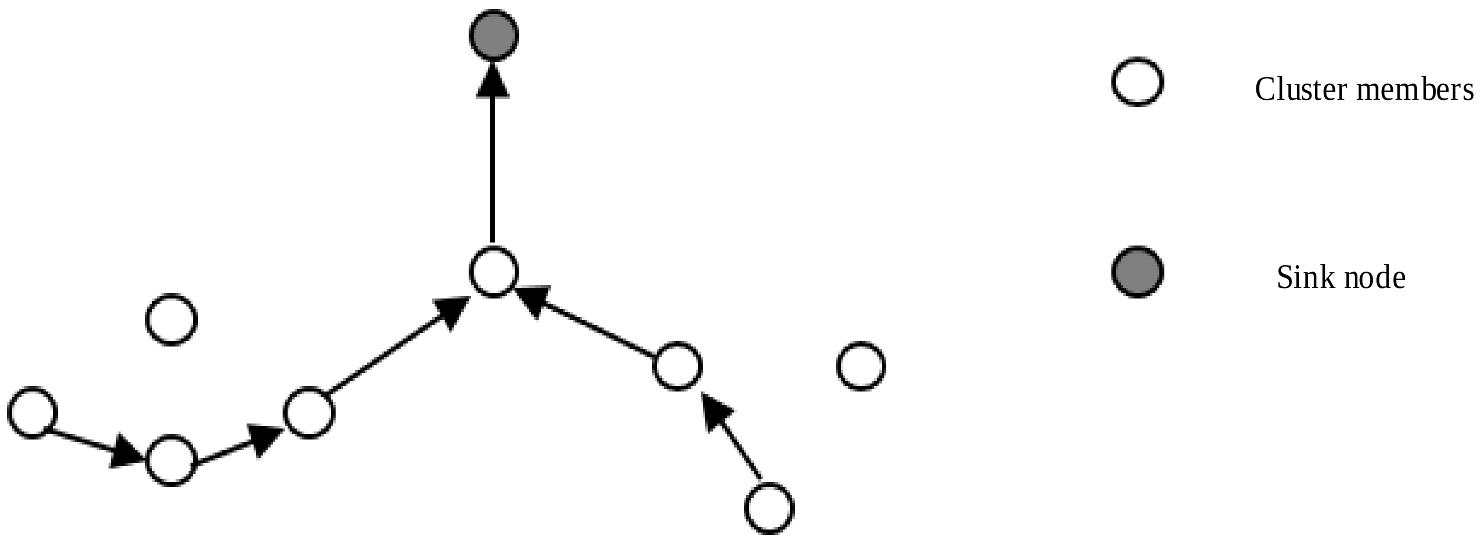
(d) Path selection and enhancement

(e) Send the data

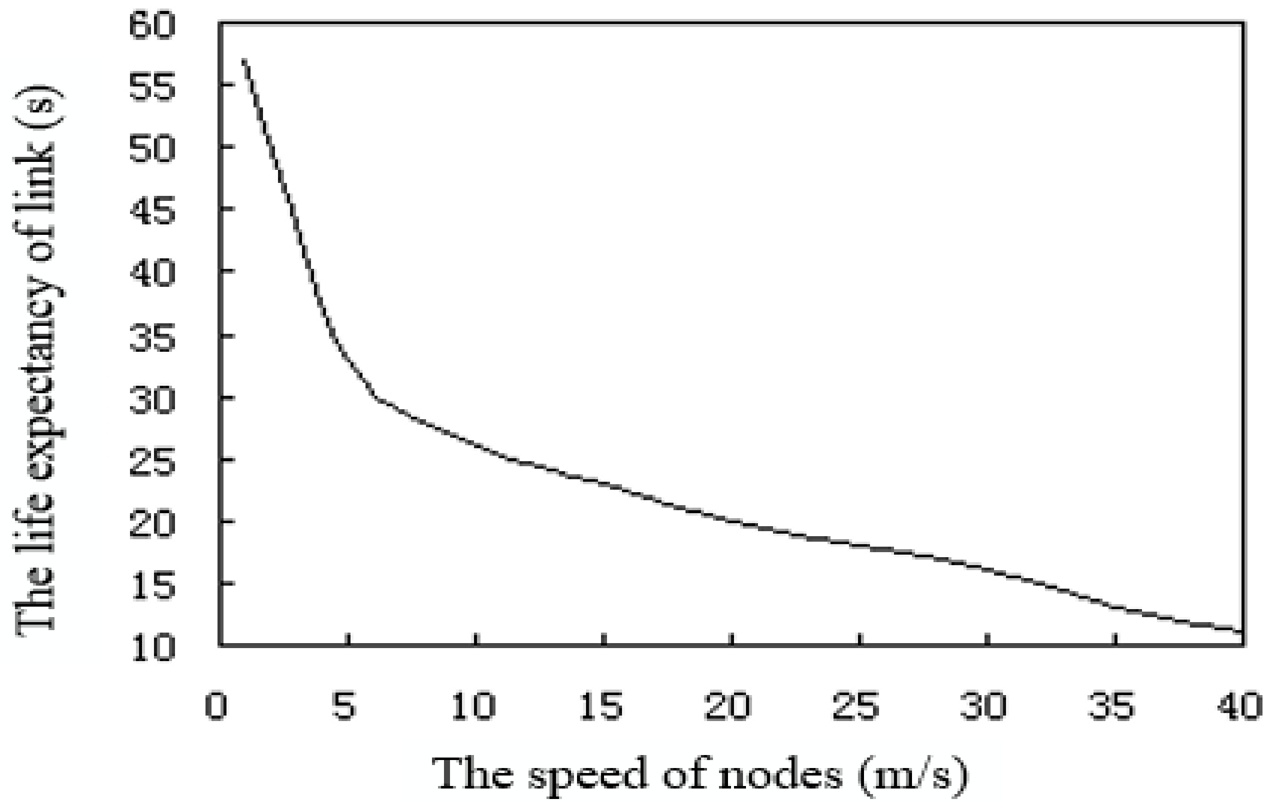
Routing protocols: hierarchical



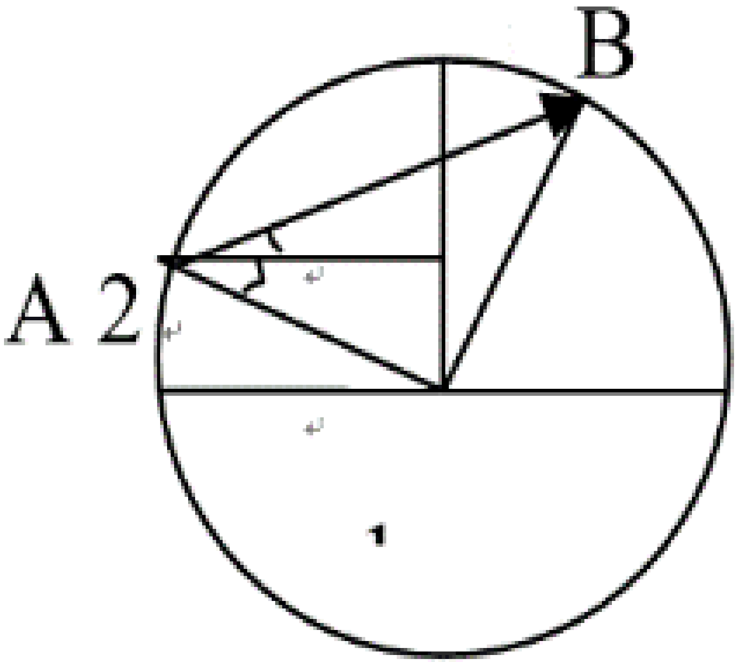
Routing protocols: chain-based



Unique challenges: speed/ energy



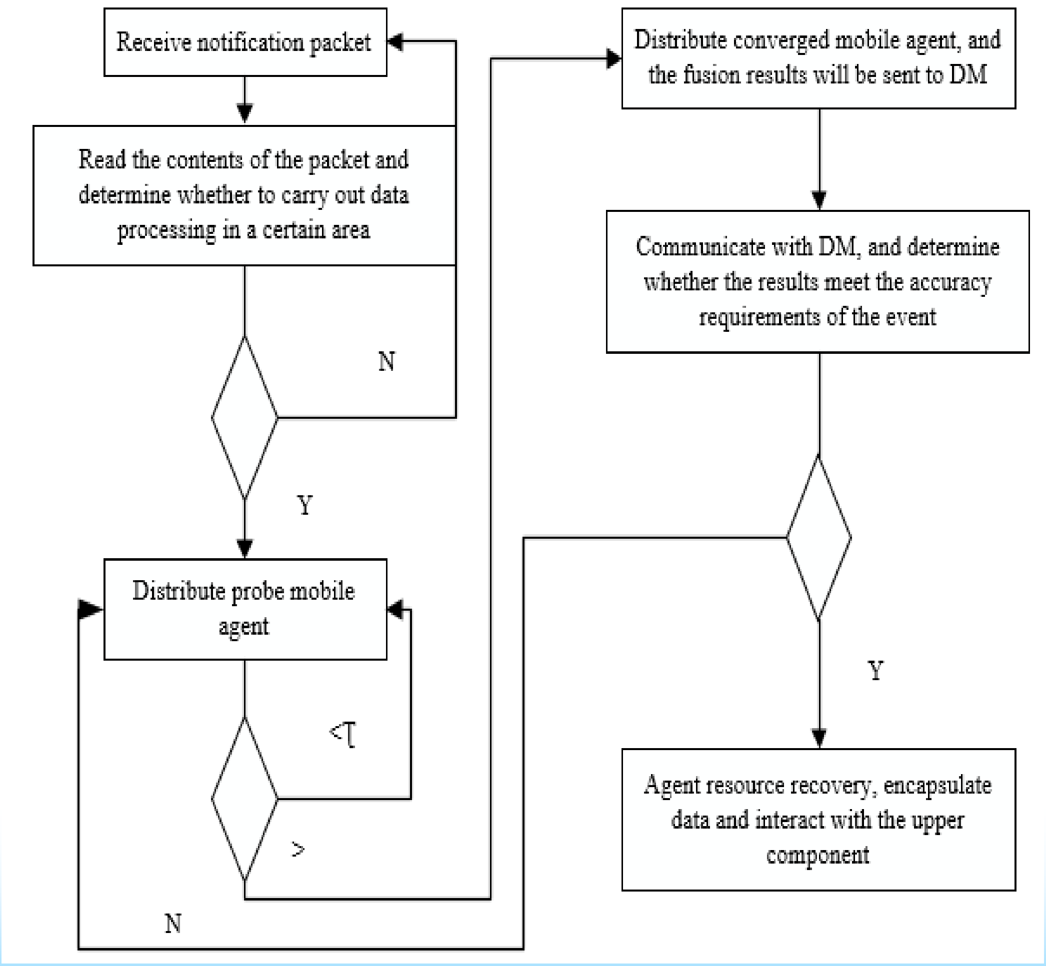
Unique challenges: mobility



Node categories

- **Sensor node** (SN): smart sensors, PDAs, laptops, mobile robots etc. communicate wirelessly.
- **Cluster head node** (CHN) : based on the needs of the application any node may become a cluster head node.
- **Regional management station** (RMS): powerful workstations communicating wirelessly with sensor nodes, and directly connected to the wired network that manages the subnet of the sensor network .

Data fusion/ integration

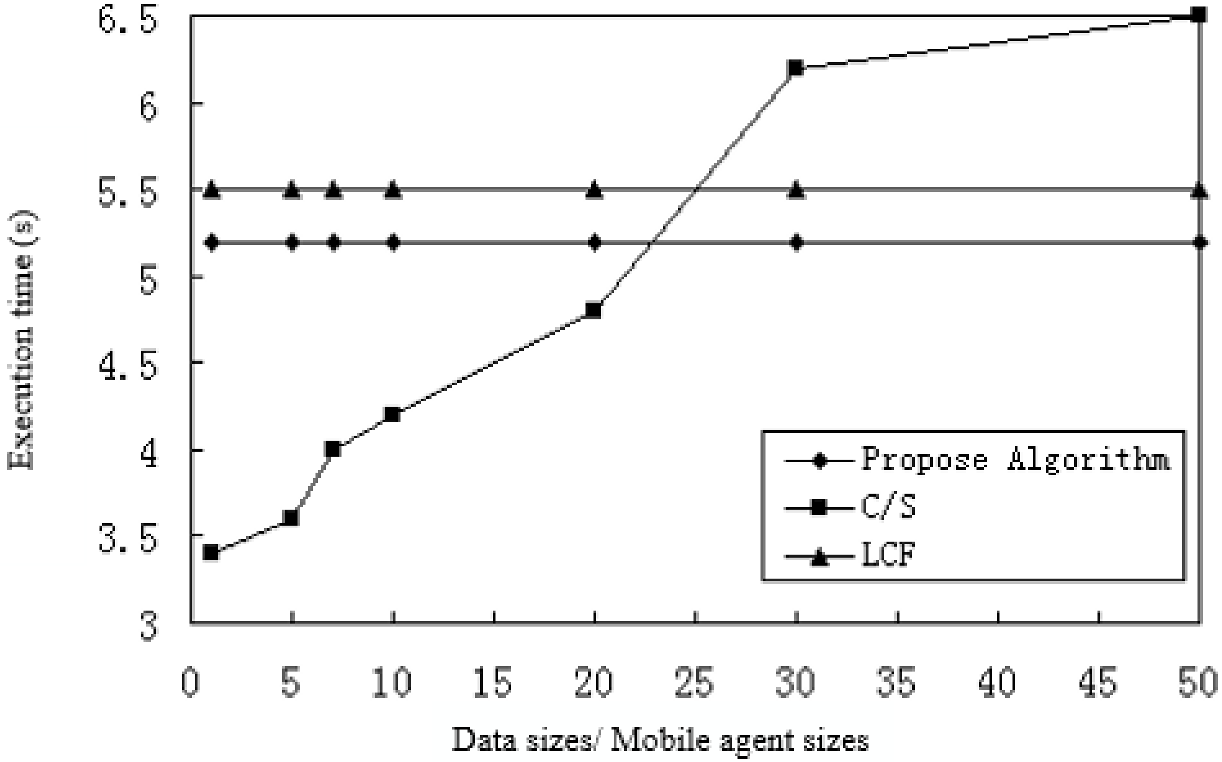


Where do supercomputers come in?

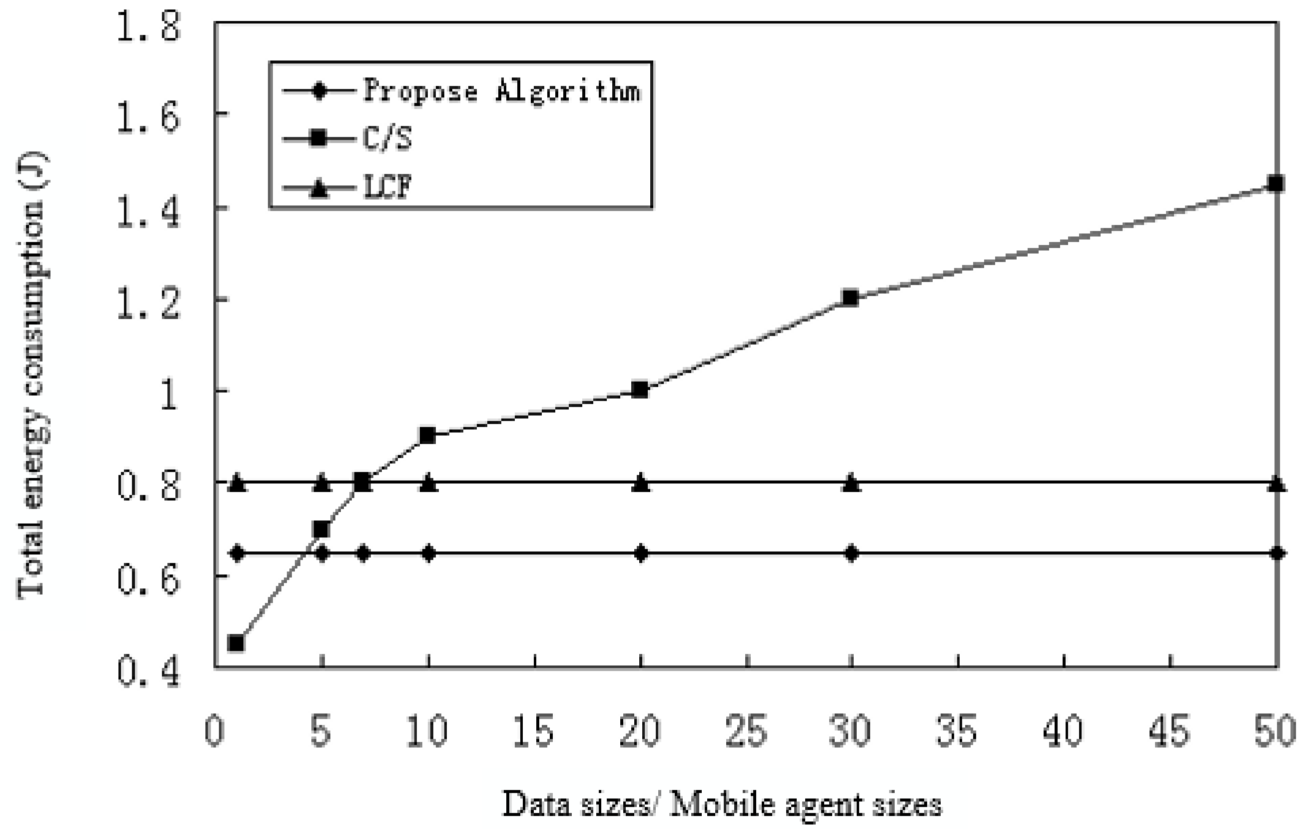
- Kind of hard to just recreate a large mobile network
- So: we simulate it on a supercomputer!

- In next slides:
  - C/S = client-server architecture
  - LCF = local closest first
  - (GCF = global closest first)

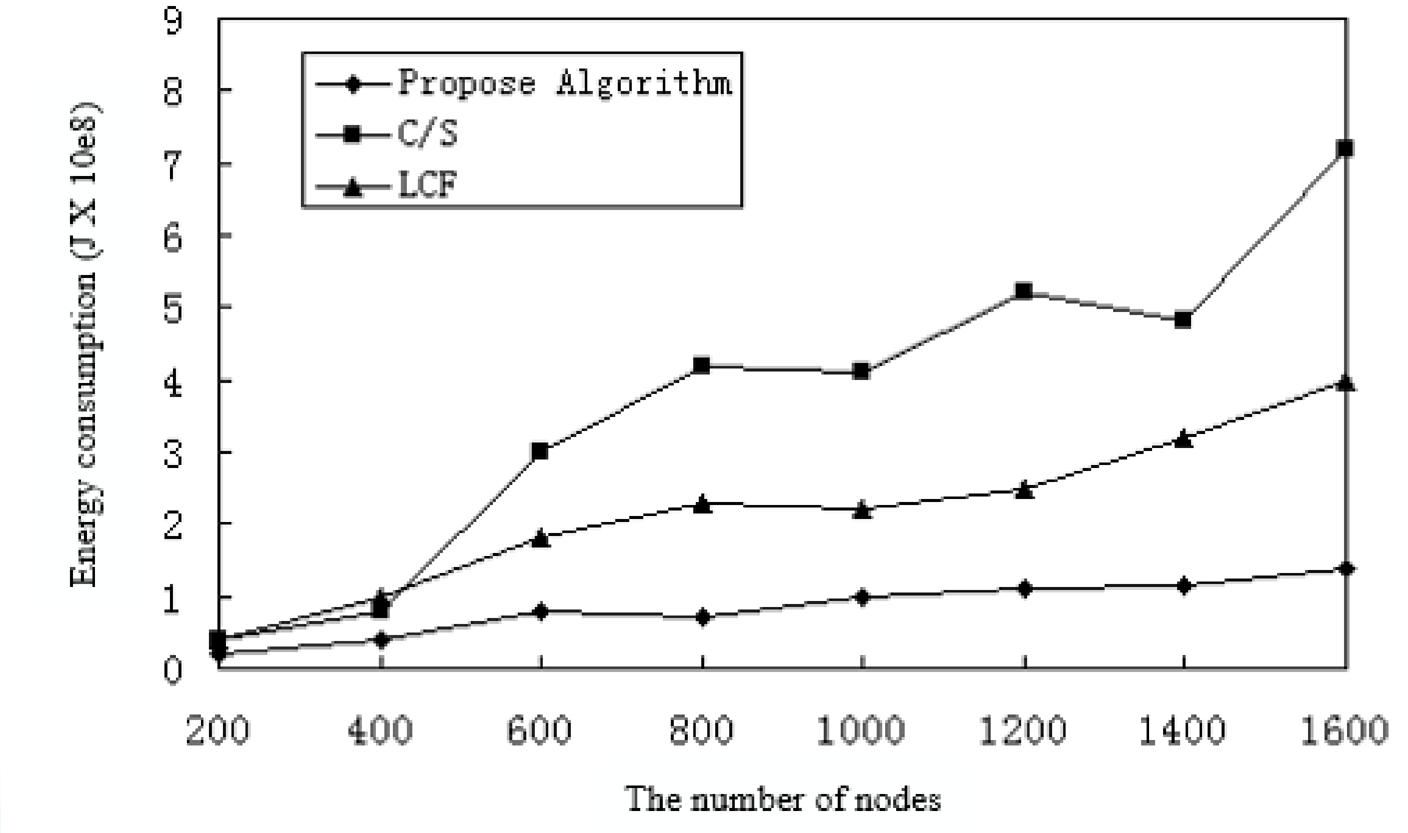
Results: Execution time



Results: Energy consumption



Results: Node density



Conclusions

- Lower energy consumption in large number of nodes
- Lower transmission time due to data fusion

History of NSU Computer Science Beowulf Computing Cluster

Gordon Shamblin and Steven Rice
Northeastern State University

History of Beowulf Cluster

- Donald Becker and Thomas Sterling designed the first Beowulf prototype in 1994 for NASA. It consisted of 16 486-DX4 processors connected by channel-bonded Ethernet. The next Beowulf clusters were built around 16 Pentium Pro (P6) 200-MHz processors connected by Fast Ethernet adapters and switches.

Needs for Cluster

- A place to put the cluster.
- Computers for the cluster.
- Networking for the cluster.
- Operating System for the cluster.
- Programming Software for the cluster.

NORTHEASTERN
STATE UNIVERSITY

A place to put the cluster.

- Rack/Table: to hold the equipment
- Power Consumption: enough to power the nodes
- Ventilation: to keep the area cool
- Physical Access: so people can work on the systems
- Internet Access: so student can access the cluster

First location for the cluster was to be in a 14’ X 10’ office on a 19” rack.

New location for the cluster is in Science 248, a 30’ X 30’ equipment room with A/C, University conectivity and power

First Cluster.

Master PC
Processor: P4 @ 1.6 GHz
Memory: 768 MB
Drives: 1-40GB, 1-20GB
Networking: 2-10/100Mbps Ethernet adapters
Number of Master = 1

Node PC
Processor: P3 @ 933 MHz
Memory: 384 MB
Drives: 20GB
Networking: 10/100Mbps Ethernet adapter
Number of nodes = 16

Spare PC
Processor: P3 @ 1 GHz & 933 MHz
Memory: 256 MB
Drives: 20GB
Networking: 10/100Mbps Ethernet adapter
Number of units: 3 @ 1GHz, 4 @ 933 MHz

Beowulf Cluster

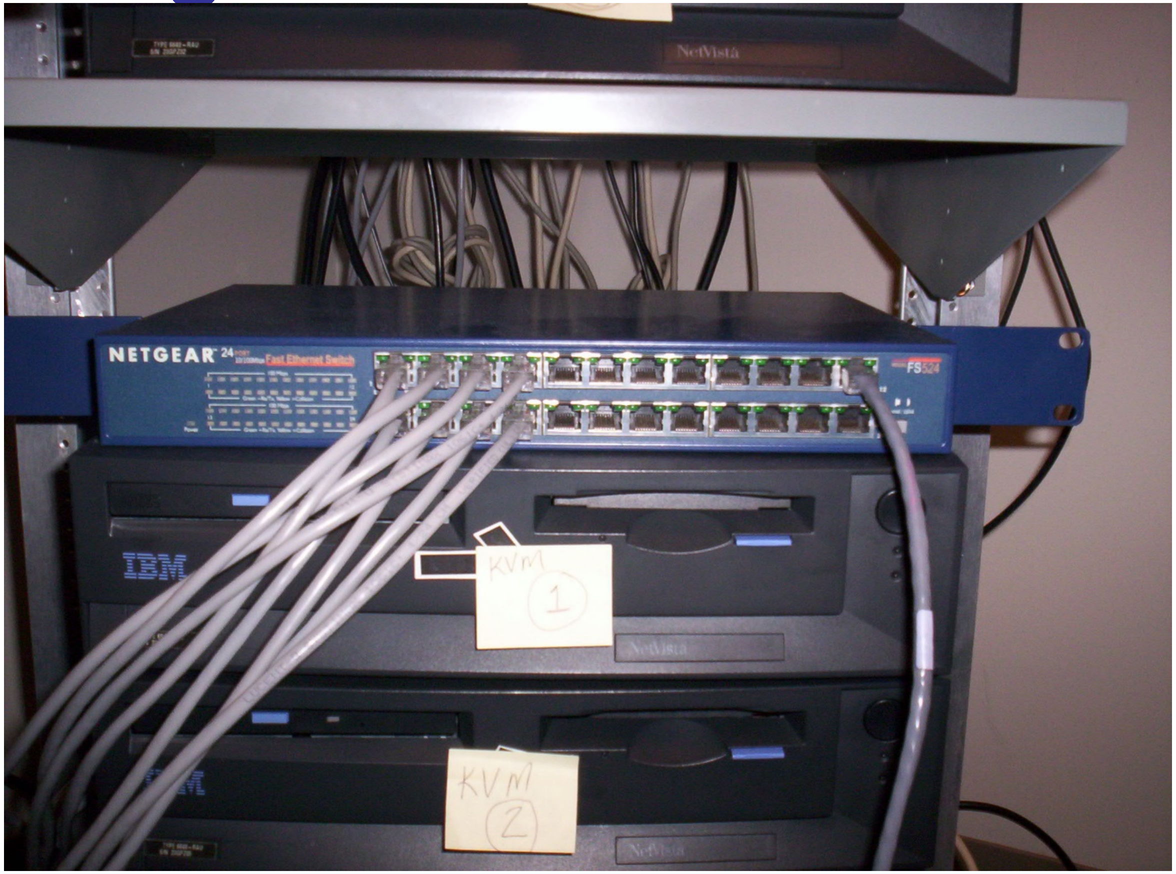
Master Node

Nodes 1 ~ 8 Nodes 9 - 16

Spares Nodes

NORTHEASTERN
STATE UNIVERSITY

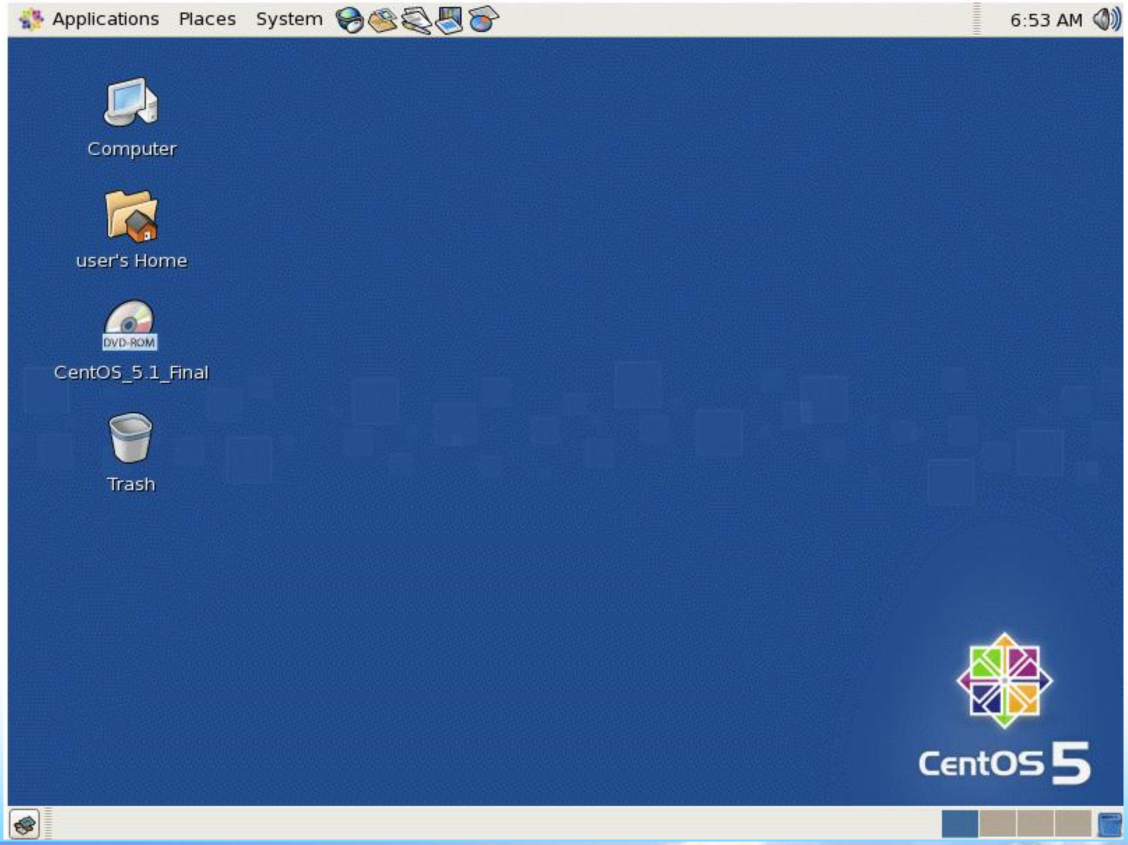
Networking for the cluster.



•24 – port Netgear FS524 10/100Mbps Switch

Operating System for the cluster.

- CentOS 5.1 Linux based on RedHat Enterprise Linux 5



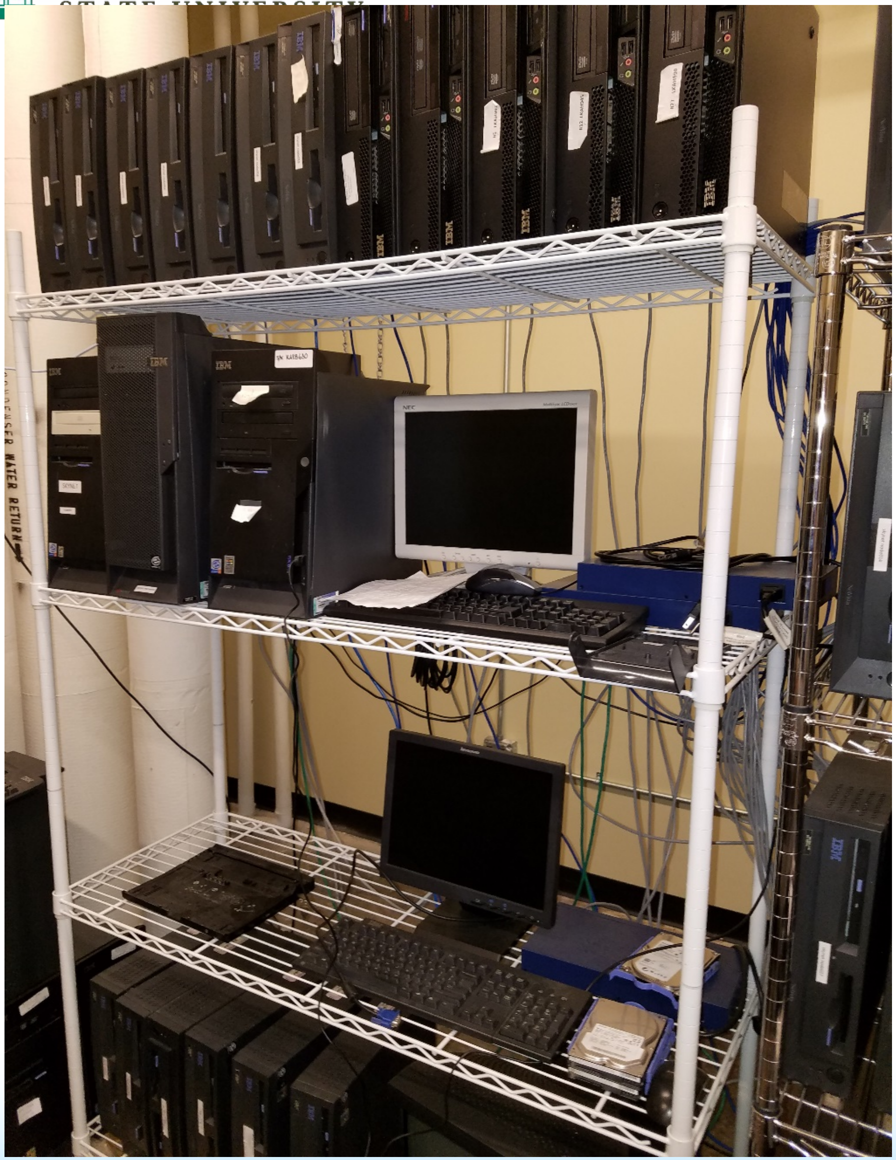
Second Cluster

This was our main Cluster for more than 6 years. As you can see it was an expansion of our first cluster. It was named: Skynet.

Master PC
Processor: P4 @ 3.0 GHz
Memory: 1 GB
Drives: 1-40GB, 1-20GB
Networking: 2-10/100Mbps
Ethernet adapters
Number of Master = 1

Node PC
Processor: P3 @ 933 MHz
Memory: 256 MB
Networking: 10/100Mbps
Ethernet adapter
Number of nodes = 64

Spare PC
Processor: P3 @ 1 GHz & 933 MHz
Memory: 256 MB
Networking: 10/100Mbps Ethernet adapter
Number of units: 6 @ 933 MHz




IBM
IBM
SKYNET


lenovo
IBM
IBM

Third Cluster

In 2014 we acquired 70 IBM PC to replace the current Skynet Cluster nodes. But with all used PC we had to get parts to upgrade the RAM to a new level of performance.

Master PC
Processor: Pentium D (Dual Core) @ 3.0 GHz
Memory: 4 GB
Drives: 1-160GB
Networking: 2-10/100Mbps Ethernet adapters
Number of Master = 1

Node PC
Processor: Pentium D (Dual Core) @ 3.0 GHz
Memory: 2 GB
Networking: 10/100Mbps Ethernet adapter
Number of nodes = 64

Spare PC
Processor: Pentium D (Dual Core) @ 3.0 GHz
Memory: 2 GB
Networking: 10/100Mbps Ethernet adapter
Number of units: 4

NORTHEASTERN
STATE UNIVERSITY

Fourth Cluster

On the way to building our third cluster, the University of Arkansas decommission it Star of Arkansas Cluster. We received 3 Dell cabinets with 64 U1 computers and the 10 Gbps networking components.

We were to house the cluster in the Webb Center building but access and power restriction delayed setup and installation. It has since been move to Science 248 with the rest of our clusters and will be reassembles there.

<u>Node PC</u>
Processor: 2 Xeon E5430(Quad Core) @ 2.66 GHz
Memory: 8 - 2GB 667 DDR2
Networking: 10 Gbps Ethernet adapter
Number of nodes = 64


OfficeMax
Computer Paper

Northeastern State University


DELL
PowerConnect
2824

Thanks for your attendance!