Fermo, September 15, 2006

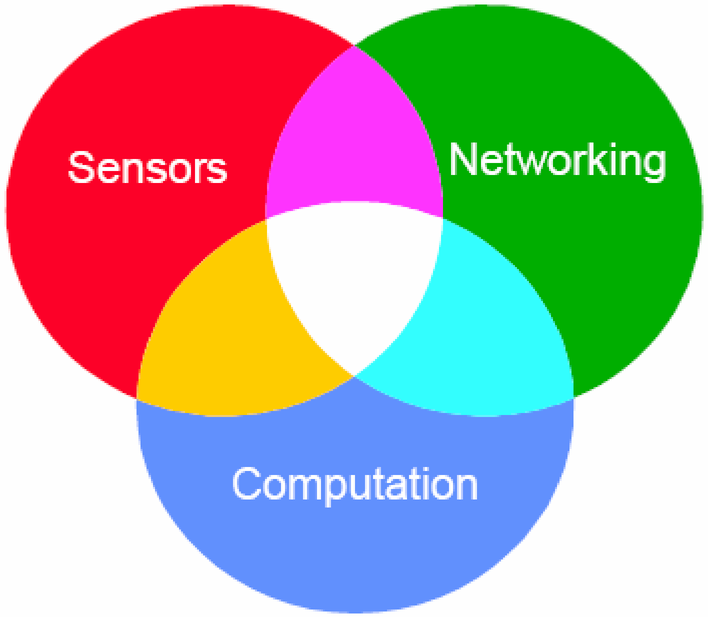
Challenges for Data Mining in Distributed Sensor Networks

VIRGINIO CANTONI
LUCA LOMBARDI
PAOLO LOMBARDI
Università di Pavia, Italy

virginio.cantoni@unipv.it
luca.lombardi@unipv.it
paolo.lombardi@jrc.it
http://vision.unipv.it

Agenda

- Preliminaries
- History of SN
- Technological issues
- Distributed data management
- Distributed hypothesis formation
- Networks of cameras
- Applications
- Evolution opportunities and challenges

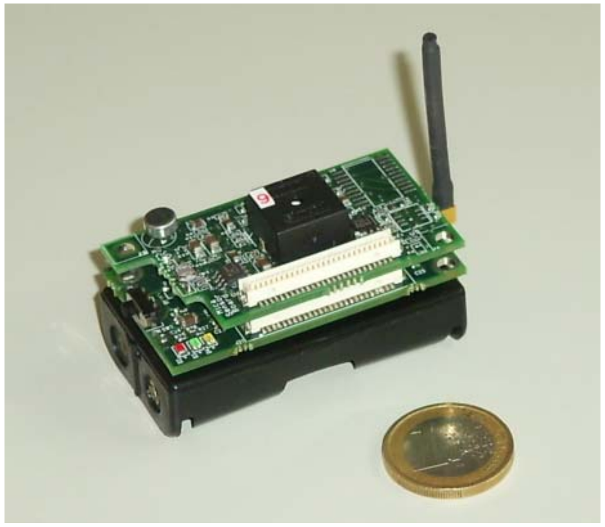


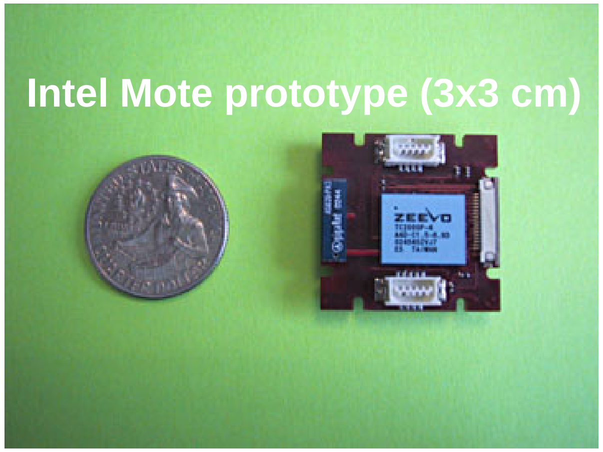
Objective of Sensor Network

- The objective of SN is knowledge-gathering and understanding such that:
  - the system can answers questions that no single device could answer;
  - each contributor has a sense of its relationship with others and acts to maximize its contribution;
  - each additional contributor potentially adds details/precision/speed to the answer.
- The devices themselves have a role in the overall process and cooperate with one another to the understanding phenomena *in situ* and in *real time.*

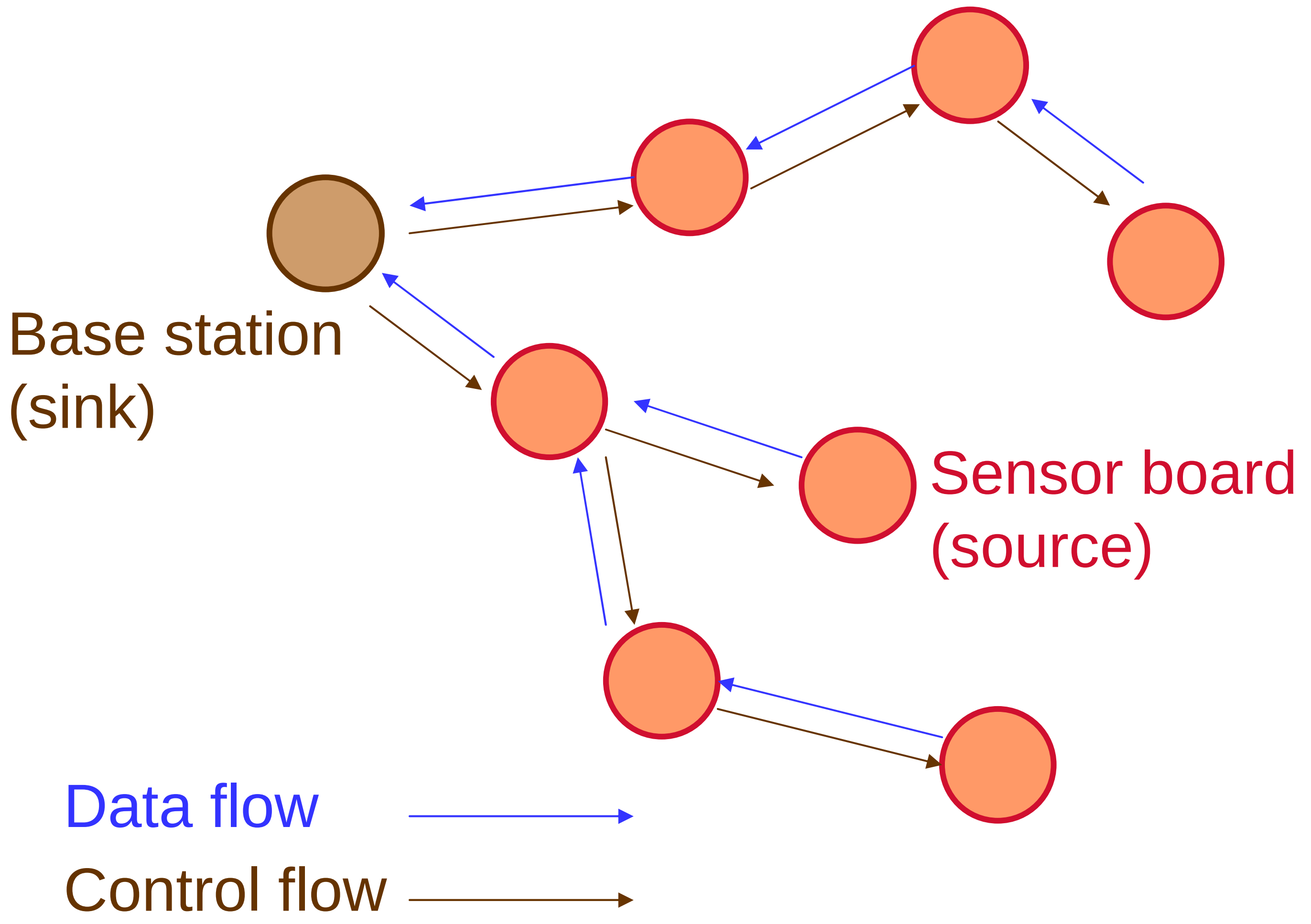
DSN,WSN, mote

- A typical sensor node is a small battery-powered board including a microprocessor, a memory, an RF transceiver and an antenna
- These elements are reduced to the limit of the lowest energy consumption and dimensions to make the sensor nodes ubiquitous and of long-lasting autonomy
- Wireless connection facilitates deployment and makes the memory remotely accessible by creating a network layer between the sensors





Typical sensor network topology



Attributes of SN

Sensors	Size:	Small / large
	Number:	Small / large
	Type:	Passive / active
	Composition or mix:	Homogeneous / heterogeneous
	Spatial coverage:	Dense / sparse
	Deployment:	fixed and planned / ad hoc
	Dynamic:	Stationary / mobile
Sensing entities of interest	Extent:	distributed / localized
	Mobility:	static / dynamic
	Nature:	Cooperative / Non cooperative
Operating environment		Benign / adverse
Communication	Networking	Wired / wireless
	Bandwidth	High / low
Processing architectures		Centralized / distributed
Energy availability		Constrained / unconstrained

Point-to-point network

- Early period (1950-1980) main characteristics:
  - generally adopt a hierarchical structure where processing occurs at consecutive levels until the information about the system reaches the user;
  - human operators play a key role in the system;
  - researches was focused on satisfying mission needs like signal processing and interpretation, tracking, and fusion.
  - SOSUS (SOund and SUrvelliance System) – NOAA (National Oceanographic and Atmospheric Administration)
  - AWACS (Airborne WArning and Control System)

From the link to the network

- Second period (1980s): distributed sensor networks; main characteristics:
  - many spatially distributed low-cost sensing nodes that collaborate with each other but operate autonomously, with information being routed to whichever node can best use it;
  - technology components included:
    - distributed software (dynamically modifiable), load balancing and fault reconfiguration;
    - high level protocol that link processes working on a common application in a resource-sharing networks;
    - processing techniques and algorithms (including self-location algorithms and knowledge based techniques);
  - very few technology components were available off-the-shelf

  - COTS (Commercial-off-the-shelf)
  - CMU Accent – Mach
  - MIT SPLICE

From the network to the communication system

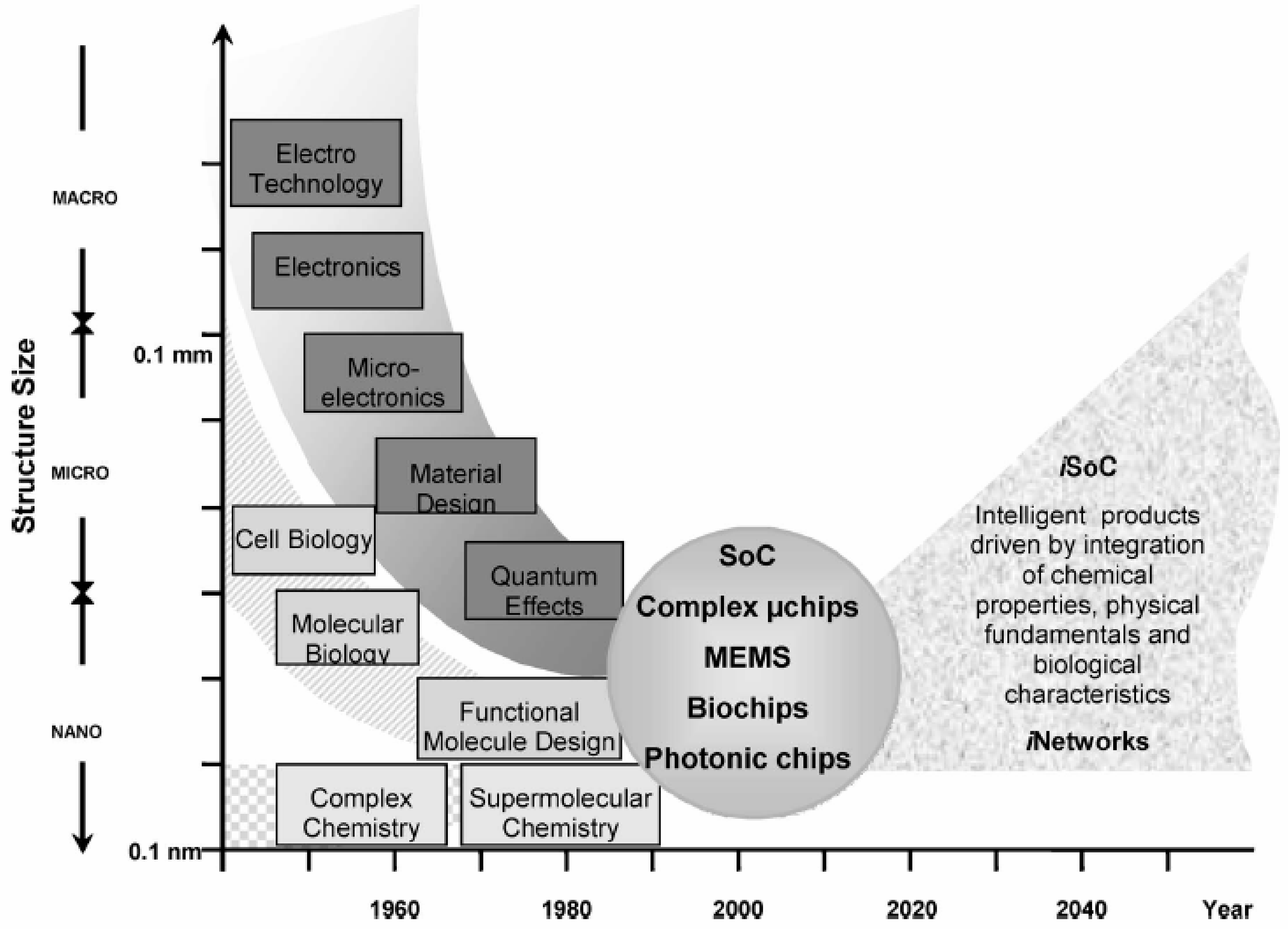
- Third period (1990s): network-centric system, main characteristics:
  - in platform-centric systems, platforms “own” specific actuators, which in turn own sensors in a fairly rigid architecture. In network-centric systems, sensors do not necessarily belong to actuators or platform, they collaborate with each other over a communication network;
  - technology components included:
    - detection and tracking performance is improved through multiple observation, with geometric and phenomenological diversity, extended detection range, and faster response time.
  - the development cost is lower by exploiting commercial network technology and common network interface.

  - CEC (Cooperative Engagement Capability)

From communication system to network information system

- Sensor network research in the 21° Century, advances in computing and communication caused a significant shift:
  - small and inexpensive sensors, wireless networking, inexpensive low-power processor allow the deployment of wireless ad hoc networks;
  - technology components included:
    - new network techniques suitable for highly dynamic environments;
    - networked information processing to extract useful, reliable and timely information from the deployed sensor network;
    - SN is interactive and programmable with dynamic tasking and query;
    - software and algorithms exploit the proximity of devices to drastically improve the accuracy of detection and tracking.
  - SensIT (Sensor Information Technology)

Bachmanns' relationship between nano, micro, and macro structure sizes



K.Eshraghian: SoC Emerging Technology, PROC IEEE, June 2006, p1197-1213

Node-on-chip

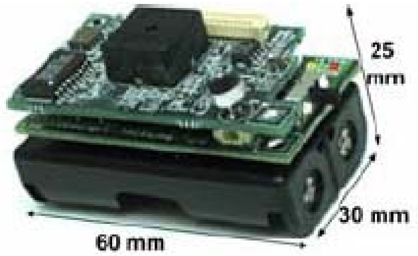
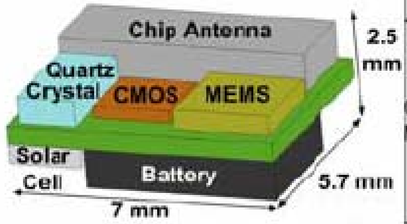
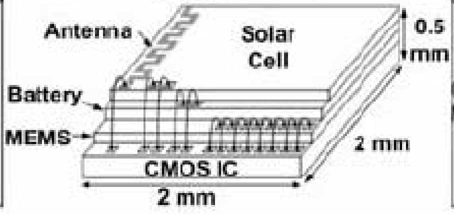
Power	Volume		Sensor(s)	Processor	Energy Harvesting	Energy Storage	RF	Antenna	Timing Reference
Highest	45,000 mm³		Off-the-shelf	Atmel Atmega 128L	None	2 AA Batteries	ChipCon CC1000	Straight Wire	Quartz Crystal
Lowest	100 mm³		CMOS - temp, light MEMS - inertial	Custom CMOS	Off-the-shelf Solar Cell	Off-the-shelf Microbattery	Custom CMOS	Dielectric Chip Antenna	Quartz Crystal
Low	2 mm³		CMOS - temp, light MEMS - inertial	Custom CMOS	Custom Solar Cell	Custom Thin-film Lithium Battery	Custom CMOS	Custom On-Chip Antenna	MEMS Resonator

Fig. 13. *A complete sensor node may be implemented with varying levels of integration. While the cost, size, and power consumption of off-the-shelf sensor nodes is far from optimal, a single-chip system may not be the most advantageous either. The most economical solution is likely to be a hybrid of integrated and assembled parts.*

B.W.Cook, S. Lanzisera, K. S. J. Pister: SoC Issues for RF Sart Dust, PROC IEEE, June 2006, p1177-1196

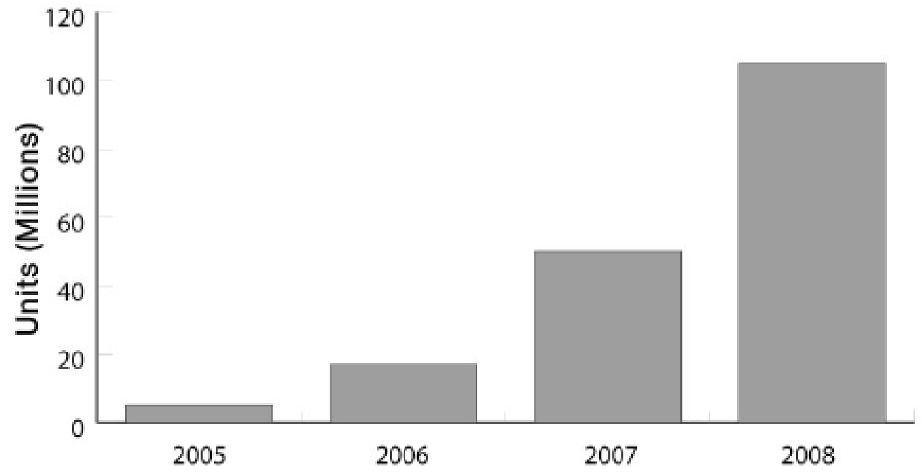
WSN market forecast

High reliability

- > 99.9% end-to-end reliability

Scalability

- Up to 10,000 motes per entire network
- Up to 250 motes per SmartMesh Manager



Crossbow

Technological Issues

Sensor nodes should not only provide hardware management for sensors, radios, I/O and memory, but also simple task coordination and networking. Basic features of the WSN:

- complexity level. Capability embedded in each single device:
  - ad-hoc software for a single hard-coded application;
  - complex operating systems able to manage multiple applications.
- power consumption. Only limited resources are available.

Example Mica2 sensor node:

Action	Energy consumed
Sample (single sensor)	1.637×10^{-6} *J*
Send (single message)	1.653×10^{-3} *J*
Listen (for 1 sec)	23.88×10^{-3} *J*
Sleep (for 1 sec)	90×10^{-6} *J*
Aggregate (max of array)	1.637×10^{-6} *J*

Technological Issues

- Memory limits. Data memory management is very critical due to its small amount:
  - a single address space both for OS and for applications, CPU works always in the same operating mode;
  - program memory is not critical because applications usually are small enough to fit into flash memory.
- Radio Frequency devices. Two different approaches are usually adopted:
  - a raw bit interface: in such case every time a bit is received the CPU is interrupted, this dramatically reduces the performance and increases the complexity of the software and power consumption;
  - a hardware packet management: in this case the CPU can be sleeping even if the radio device is sending or receiving a packet.

Technological Issues

- Communication medium. The Media Access Control (MAC) should satisfy several constraints:
  - fairness: the MAC layer should be fair, that is every device in the network should have the same amount of shared resources;
  - latency: the message latency is the time a message takes to route from source to destination. A short latency is desirable, but a high latency reduces the power consumption;
  - positioning: the MAC software must be able to detect device position and organize the topology of the network;
  - heterogeneity: the MAC layer must be able to manage networks with a large range of devices;
  - scalability: some applications need only a few nodes in a small room, others thousands of sensors in large areas;
  - self-organizing: wireless sensor networks usually operate in unstructured environments and without human monitoring.

Self- properties*

- A system is self-organizing *"if a collection of units coordinate with each other to form a system that adapts to achieve a goal more efficiently“,* in particular these connotations are addressed:
  - self-configuration – the system determines the arrangement of its constituent parts;
  - self-localization - randomly distributed nodes determine a global coherent positioning;
  - self-optimization - the system monitor and control resources to ensure optimal functioning;
  - self-awareness - the system has a sense of its own behavior, a knowledge of its components, its current status, its goal, and its connections to other systems;
  - self-healing - the can repair itself, or at least take actions so that faults that appear in the system do not hinder the application currently running;
  - self-tuning, self-maintenance, self-(re-)production, self-regulation, self-steering, self-reference, self-empowerment, self-protection.

A. Pilpre':Self- Properties of Multi Sensing Entities in Smart Environments, June2005, MIT Thesis*

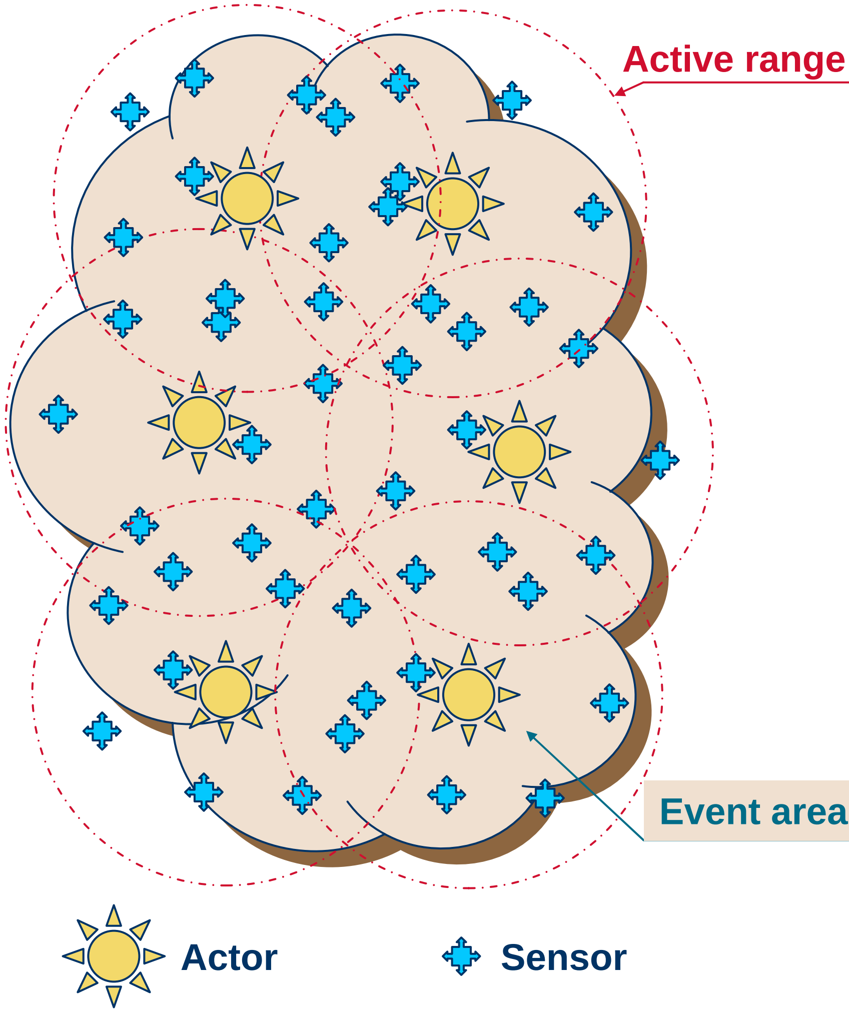
Technological Issues

- Operating systems: an open-source operating system designed for WSN application is TinyOS. Basic characteristics of TinyOS:
    - the capability of managing a distributed application;
    - very rigid constraints on the amount of available memory;
    - adoption of a single stack, task cannot be interrupted before its natural end;
    - application is made up of a set of “components”, each one is coded with a well defined interface both for input and output;
    - when a component requires a particular service declares un “event” implemented in the source code;
    - an application is mapped in a directed graph in which components correspond to nodes and commands and events to edges;
    - each task must take care of all system resources and their management.

Wireless Sensor and Actor Networks

WSANs objectives:

- provide real-time services with given delay bounds, according to application constraints;
- ensure an efficient communication among sensors and actors;
- ensure ordering between the different events when they are reported to the actors (ordered distribution);
- provide synchronization among different sensors reporting the same event to multiple or same actor in order to facilitate a one-time response in the entire region (S/A coordination);
- track and report the sensed event to a different set of actors exploiting the coordination between actor nodes in order to maximize their overall task performance (A/A coordination).



Data management

Data management is the task of collecting data from sensors, storing data in the network, and efficiently deliver data to the users. Basic features of the DM:

- trade-offs energy-efficiency: low energy consumption vs. timeliness of data delivery, accuracy, network latency, and bandwidth used;
- flexibility: data storage and distribution strategies must be adaptive to various topologies and to topology changes;
- robustness: data management must be robust or at least redundant, data replication and route redundancy provide robustness to data loss;
- locality: WSN promotes local management and processing strategies vs. more traditional global approaches.

Data management

The strategy of data management strongly depends on the metaphor used to understand a sensor network. The proposed paradigms are:

- distributed database: data gathering is formulated as a database retrieval problem:
  - sensors are seen as distributed storage points, which can be addressed on demand of the user;
  - sensor programming coincides with query dissemination and processing.
- agent system: sensors are agents interacting according to multi agent paradigms;
  - macro-programming is based on modifying global parameters of cumulative behaviors.

Data management

The database view of sensor networks has so far attracted more attention. Sensors are queried in a SQL-like fashion. Reported query types are several:

- simple one-shot queries regard the value of a sensor at a certain time (in the past, present, or future);
- event-based queries are triggered by the verification of an event specified in the query;
- aggregate queries imply a form of data aggregation, i.e. a processing of data resulting in a more compact (lossy) representation.
  - temporal aggregation;
  - region-based operations;
  - logical expressions on outputs of multiple sensors (sometimes called multi-dimensional queries).

Query processing techniques

Technique	Summary
Event-based query	Avoid polling overhead
Lifetime query	Satisfy user-specified longevity constraints
Interleaving acquisition/predicates	Avoid unnecessary sampling costs in selection queries
Exemplary aggregate pushdown	Avoid unnecessary sampling costs in aggregate queries
Event batching	Avoid execution costs when a number of event queries fire
SRT	Avoid query dissemination costs or the inclusion of unneeded nodes in queries with predicates over constants attributes
Communication scheduling	Disable node’s processors and radios during times of inactivity
Data prioritization	Choose most important samples to deliver according to a user-specifies priorization function
Snooping	Avoid unnecessary transmissions during aggregate queries
Rate adaptation	Intentionally drop tuples to avoid saturating the radio channel, allowing most important tuples to be delivered

Data management

Fundamental services of the data management middleware, as seen in the database paradigm, are:

- data storage: distributed data storage can be delegated to the base station (sink), which usually has higher storage capacity (external storage), or it can take advantage of memory localized on sensor boards (local storage);
- query/control dissemination: routing of queries and control messages from the sink to the nodes, and the backward route of data can be made energy-efficient by smart strategies;
- data aggregation: aggregation strategies reduce the number of transmitted data, and thus improve energy-efficiency (fewer packets sent, shorter listen time), pursue an efficient use of bandwidth, and prove more scalable.

Data management

Data storage and query/control dissemination are sometimes identified together as the “data-gathering” service. Chen and Hou have proposed a taxonomy:

- data-gathering algorithms are classified along three dimensions:
  - data storage method (external, local, data-centric);
  - direction of query diffusion (double-pull, single-pull, push);
  - structure of dissemination (tree, grid, cluster, chain).
- data-aggregation algorithms are classified along three dimensions:
  - aggregation functions (SQL-like, redundancy suppression, parameter estimation);
  - aggregator architecture (i.e. placement of aggregators);
  - trade-off of resources (accuracy, latency, bandwidth).

Data management

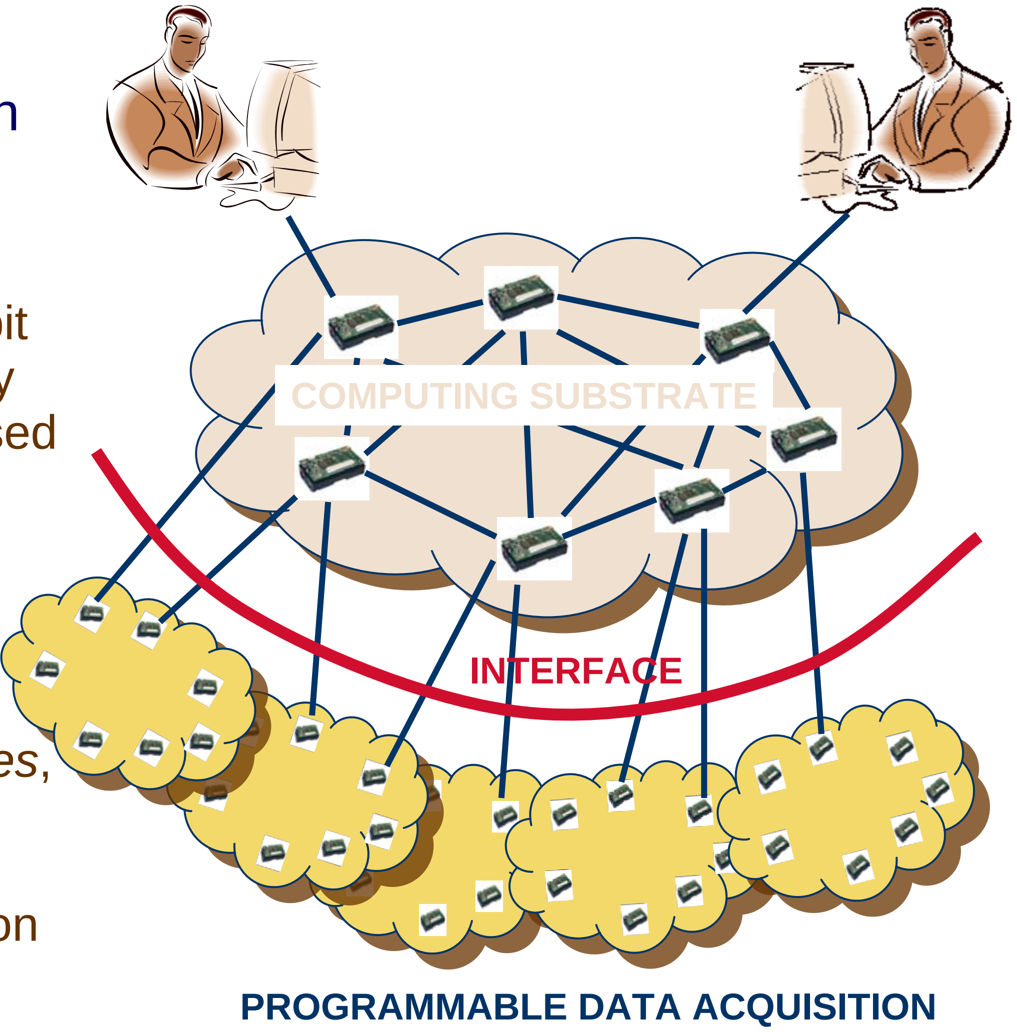
Mobile nodes pose further challenges and open possibilities for more complex in-route data mining and fusion:

- mobile nodes are often associated with more technologically endowed platforms, with engines and higher energy resources than other nodes;
- mobile nodes cause the topology to change continuously, so that geographic data storage and topology-dependent aggregation must be made adaptive;
- mobile nodes can provide faster reaction (lower latency) and higher energy-efficiency by reducing the number of transmission needed to reach a sink.

Tenet: an architecture for tiered SN

The Tenet tiered architecture prescribes a functional separation between motes and masters:

- the upper-tier, consisting of 32-bit nodes, *masters*, is free of energy constraints and provides increased network and computational capacity, enabling large-scale deployments;
- the lower-tier, composed of *motes*, contains sensing and actuation functionality and enables infrastructure-less instrumentation of physical spaces and artifacts.



Tenet: an architecture for tiered SN

The Tenet is based on the following design principles:

- addressability. Any master can communicate with any mote or master, as long as there is (possibly multi-hop) physical-layer connectivity between them. Any mote can communicate with at least one master, unless no master is reachable;
- asymmetric task communication. Any and all communication from a master to a mote takes the form of a task. Any and all communication from a mote is a response to a task; motes cannot initiate tasks themselves;
- task library. Motes provide a limited library of generic functionality, such as timers, sensors, simple thresholds, data compression, and data transformation. Each task activates a simple subset of this functionality, and may also command actuators such as robots, cameras etc., and invoke node management functions like setting device parameters or returning node statistics;
- robustness and manageability. Finally, a Tenet must be robust to node and link failure, and must provide users with insight into network problems as well as allow automated response to those problems.

Hypothesis formation

- The naive strategy of collecting all data in a central computing node with a high computational power does not optimize the use of energy-costly transmissions.
- “Decentralized” and “distributed” algorithms are two common solutions:
  - decentralized algorithms assume that each node is linked to every other node in the network. As a consequence the total number of links is $O(n^2)$ (where n is the number of sensors), so decentralized algorithms scaling is difficult;
  - distributed algorithms are based on networks in which the number of links is $O(n)$ or at most $O(n \log(n))$. A fully connection scheme is therefore not allowed in distributed computing.

Consensus Filter

Consensus filter consists in the fusion of a set of sensor measurements.

In many applications the problem consists in the data fusion of a sensor network of size n where each node measures a signal corrupted:

$$u_i(t) = r(t) + v_i(t),\ i = 1, \dots, n$$

where r(t) is usually a low frequency signal while v(t) is a high-frequency Gaussian noise.

- Duarte-Hu propose a distance-based fusion algorithm: sensors far from the target have a significantly lower chance to produce a reliable value;
- Rabbat et al. propose a solution in which each node is visited only once because the network identifies a path connecting all sensor. A single process flushing from the start node to the ending one allows a global average of a parameter.

Consensus Filter

- Willet et al. propose an approach based on a central computation and coordination node: in order to obtain a long life network the adaptive sampling is based on two steps:
  - in the first step (preview step) only a subset of the sensors are active and produce a raw estimate of the environment parameters;
  - the second step (refinement step) is based on the previous results and activates the most promising nodes to refine the previous, partial results.

  By maintaining the number of active sensors as small as possible a low energy consumption is obtained;
- Olfati-Saber et al. propose a scalable distributed Kalman filtering scheme that can track relatively fast varying signals.

Distributed tracking in WSN

In SN sensors should share data to exploit sensor data fusion without sending data requests to and collecting data from all sensors, thus overloading the network

- Zhao *et al.* propose a solution to determine dynamically which sensor is most appropriate to perform the sensing, what needs to be sensed, and to whom to communicate the information.
  - Each sensor computes the predicted information utility of a piece of nonlocal sensor data and uses this measure to determine from which sensor to request data.
- Sensor cooperation and data association is needed in tracking multiple targets that are close to each other. An approximate approach for identity management is proposed by Shin et al.

Networks of cameras

- Networks of cameras have several characteristics that affect the system architecture:
  - the sensors may have the ability to pan, tilt, zoom, or even relocate;
  - the raw data is potentially of overwhelming size;
  - the algorithms are complex, multi-dimensional, nonlinear and often consist of independent followed by integrative processing;
  - the sensors typically do not cover an area as densely as simpler sensing devices would.
- Long-range transmission of raw data would cost order of magnitude more power than local processing.
- A variety of approaches have been proposed that keep the data at or near the sensors.

V. M. Bove Jr.and J. Mallet: Eye Society Project

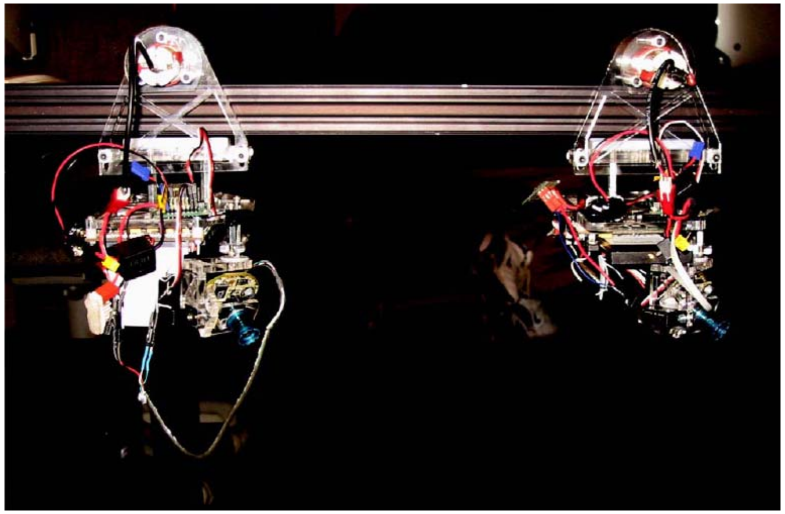
Networks of cameras

- The local processing performed on a camera will typically result in a significant reduction of the amount of data transmitted among collaborating cameras:
  - event-of-interest detection (transmit only if the event is happening);
  - region-of interest identification (transmit only a spatial region in which the sought-for feature is visible);
  - feature or metadata extraction (edge maps, feature vector, etc);
  - independent data compression or compression conditioned upon knowledge of other sensors (e.g. send only the differences);
  - high-level model building, where the model is more compact than the set of individual 2D views.

V. M. Bove Jr.and J. Mallet: Eye Society Project

'Eye society'

- The "society" consists of small, cheap, autonomous wireless mobile cameras each controlled by an embedded processor.
- Each camera can independently pan and tilt and move along an overhead lighting track, and is in constant communication with its fellow cameras, and is independently capable of analyzing its own data, and sharing information on what it sees with its fellow robots.



- The fundamental principles behind the Eye Society project are:
  - It is possible to put sufficient processing into an information capture device such that a central processing server is not needed for many sophisticated tasks.
  - A user or programmer does not need to think in terms of an identifiable individual device but of the overall system.

V. M. Bove Jr.and J. Mallet: Eye Society Project

The role of groups in CN

Among autonomous elements, group creation becomes the problem of providing all the information required for individuals to decide whether or not they can usefully contribute to the group. The group protocol includes in the group creation announcement information on the task to be performed, and a set of task-specific joining requirements like:

- creating a group and advertising its membership criteria;
- joining or leaving a group;
- sending and receiving messages within a group;
- modifying the membership criteria;
- re-evaluating whether to remain in the group (whenever another enters or leaves, or the criteria are modified, the group members are notified of the change and may need to re-evaluate their own membership);
- terminating a group.

Jacky Mallet':Self- The role of group in smart camera network", February 2006, MIT Thesis*

Applications

- Air traffic control
- Building and structure monitoring
- Distributed robotics
- Entertainment
- Environment monitoring
- Industrial and manufacturing automation
- Intelligent Health-Care Network - *i*HcN
- Military sensing
- Physical security
- Traffic surveillance
- Urban sensing
- Video surveillance
- Wearable sensors

Opportunities and challenges

- Distributed sensor networks are indeed an attractive technology, even if runtime execution cannot be achieved for complex pattern recognition because of the technological limits of the program/stack memory and the battery life.
- Of course, effective pattern recognition and data mining can be implemented on the central base station, where the computational power is not generally constrained. However, the computational capabilities of nodes will surely grow in the future, and the overall network performance would gain from distributed algorithms.
- The future scenarios of a “sentient world” or of an “internet of things” are future realistic scenarios standing good chances to be reality in ten years from now. All authors also pointed to the technological challenges posed by the increasing, and necessary miniaturization.
- Besides the hardware-related challenges, all such data will have to be stored, compressed, and analyzed. Sensor networks may provide in the future very heterogeneous data: image, sound, distance, acceleration, and maybe smell or even others.
- Pattern recognition, data mining, and data fusion will be key issues both for summarizing the data into “events”, and for elaborating reactive decisions.