Workflow Roles in Processing Large Undersea Videos for F4K

Gaya Nadarajan
gaya.n@ed.ac.uk

www.fish4knowledge.org

F4K

- URL www.fish4knowledge.org
- Three-year E.C.-funded project (Oct’10)
- Partners:
  - Edinburgh (Image Processing - IPAB, Workflow - CISA)
  - CWI, Netherlands (User Interface, Query & Analysis)
  - U. Catania, Italy (Image Processing)
  - NCHC Taiwan (Data, HPC)

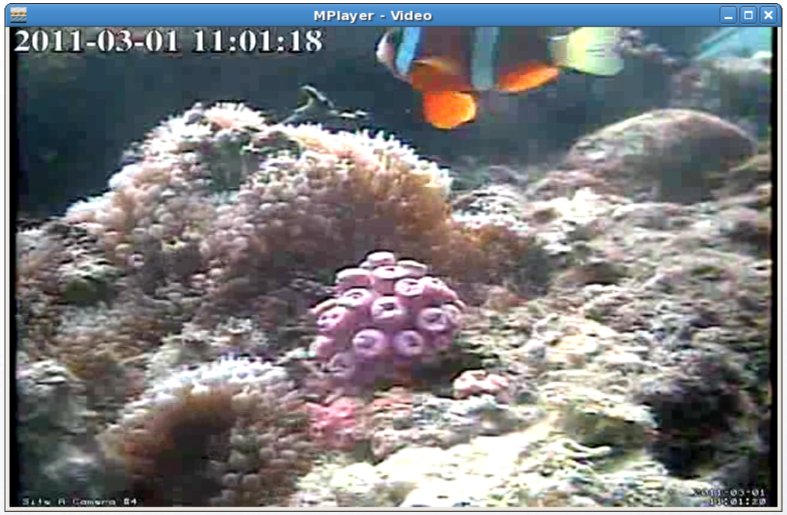
F4K Aims

- Automatic storage and extraction of information of observed marine life from database
- Provide interfaces for non-expert users (marine scientists) to query on

F4K Facts

- >3000 species to be identified
- Continuous collection: 24-25KB videos & counting, 8-30fps, 20-50MB each, 20GB per hour, 100TB per year
- 10 static cameras in 4 sites, resolution 680x480/1280x720 (CCTV/HDV): http://homepages.inf.ed.ac.uk/rbf/Fish4Knowledge/resources.htm#

Image captures



20 Questions

1. How many species appears and their abundance and body size in day and night including sunrise and sunset period.
2. How many species appears and their abundance and body size in certain period of time (day, week, month, season or year). Species composition change within one period.
3. Give the rank of above species, i.e., listed according to their abundance or dominance. How many percent are dominant (abundant), common, occasional and rare species.
4. Fish color pattern change and their behavior in the night for diurnal fish and vice versa for nocturnal fishes.
5. Fish activity with one day (24 hours).
6. Feeding, predator-prey, territorial, reproduction (mating, spawning or nursing) or other social or interaction behavior of various species.
7. Growth rate of certain species for a certain colony or group of fishes.

20 Questions

8. Population size change for certain species with one period of time.
9. The relationship of above population size change or species composition change with environmental factors, such as turbidity, current velocity, water temperature, salinity, typhoon, surge or wave, pollution or other human impact or disturbance.
10. Immigration or emigration rate of one group of fishes inside one monitoring station or one coral head.
11. Solitary, pairing or schooling behavior of fishes.
12. Settle down time or recruitment season, body size and abundance for various fishes.
13. In certain area or geographical region, how many species could be identified or recognized easily and how many species are difficult. The most important diagnostic character to distinguish some similar or sibling species.

20 Questions

14. Association among different fish species or fish-invertebrates.
15. Short term, mid-term or long term fish assemblage fluctuation at one monitoring station or comparison between experimental and control (MPA) station.
16. Comparison of the different study result between using diving observation or underwater real time video monitoring techniques. Or the advantage and disadvantage of using this new technique.
17. The difference of using different camera lens and their angle width.
18. Is it possible to do the same monitoring in the evening time.
19. How to clean the lens and solve the biofouling problem.
20. Hardware and information technique problem and the possible improvement based on current technology development and how much cost they are.

Points to Consider

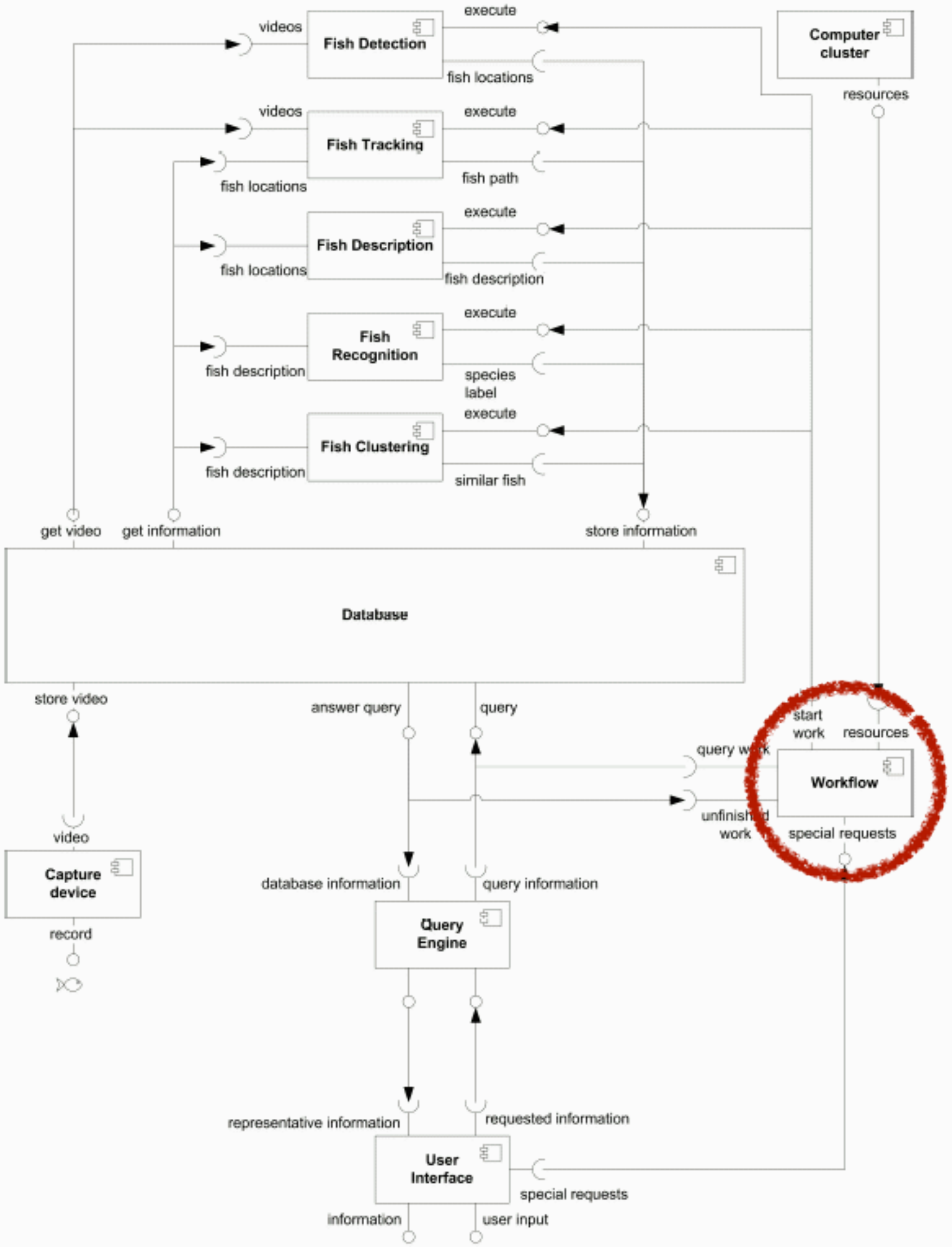
- Streams *vs.* files
- Remote access *vs.* local vs. both
- Filesystem *vs.* database
- Compute-intensive *vs.* data-movement intensive

F4K Dynamic Aspects

- Multiple streams from multiple (different) cameras
- Varying frame rate (8fps-30fps)
- Different resolutions 680x480/1280x720 (CCTV/HDV)
- Different qualities due to different resolutions and cameras
- Dynamic query formulation and solution composition & execution
- User feedback in improving solutions/overall performance
- Discovery of new knowledge can affect/improve queries

Workflow Roles

- Deal with special queries
- Manage and schedule jobs/processes to cluster optimally
- Deal with live data streams


videos
Fish Detection
execute
fish locations
Computer cluster
resources
videos
Fish Tracking
execute
fish locations
fish path
execute
Fish Description
fish locations
fish description
execute
Fish Recognition
fish description
species label
execute
Fish Clustering
fish description
similar fish
get video
get information
store information
Database
store video
answer query
query
start work
resources
query work
Workflow
unfinished work
special requests
video
Capture device
record
database information
query information
Query Engine
representative information
requested information
User Interface
special requests
information
user input

Considerations

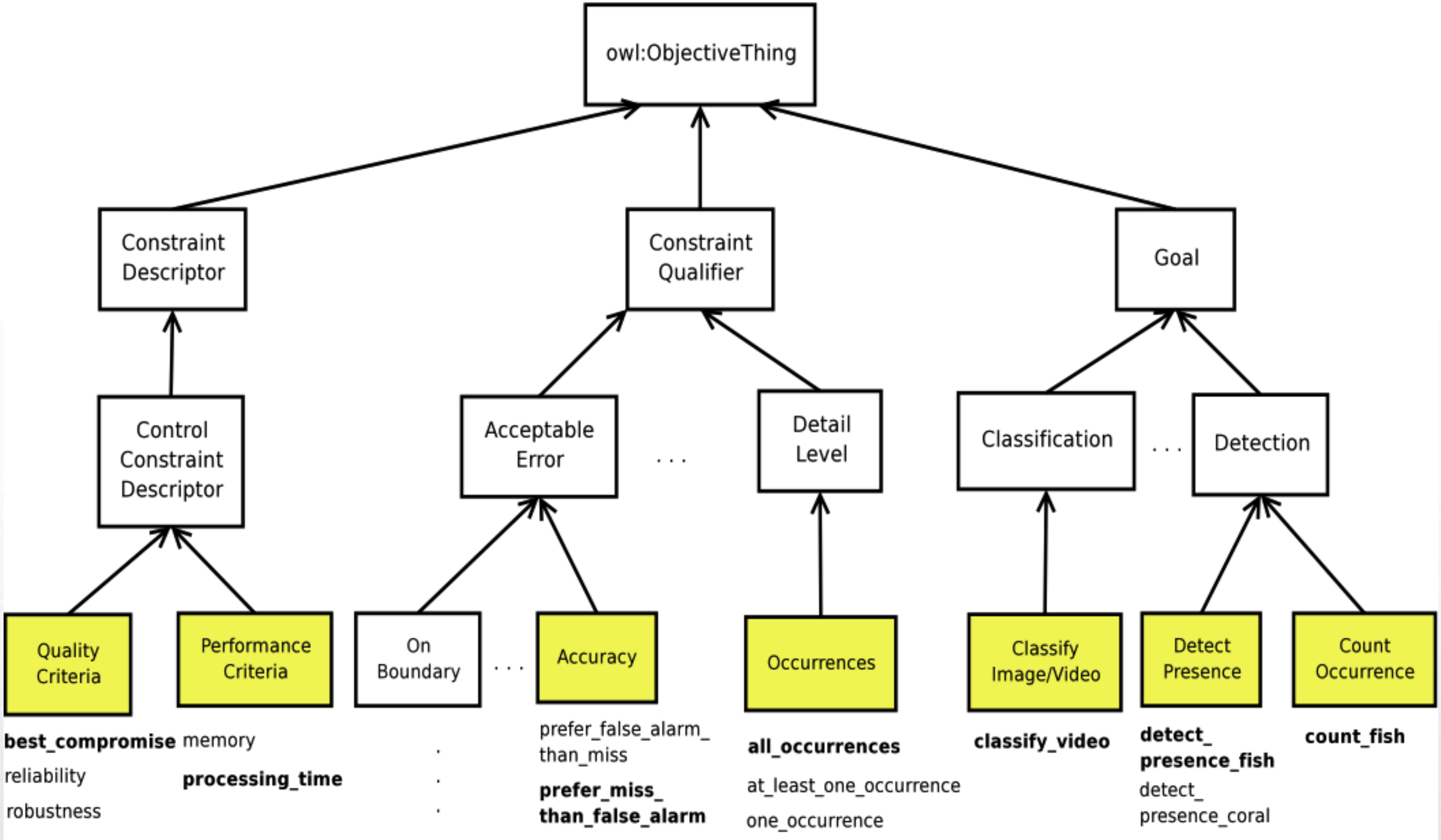
- Distributed and/or parallel processing
- Batch processing
- Data streaming instead of job-scheduling to minimise volume of intermediate data (e.g. ADMIRE)

1. Special Queries

- Queries need to be formalised into machine readable VIP goals and solutions.
- Done via ontology and planning based workflow composition mechanism (see PhD Thesis 2010)
- Assumes that multiple executables exist for VIP tasks
- Can deal with user preferences (e.g. processing time, accuracy of result) and can read in descriptions of video (e.g. brightness, clearness levels) to help with context-sensitive selection of optimal VIP modules

Query Management

- Can 20 questions be formulated using existing goal ontology? If yes, then VIP is done for free. Otherwise, enrich ontologies, or interactively formulate query
- Should workflow deal with queries directly in order to determine offline or online processing? Could do.



owl:ObjectiveThing

Constraint Descriptor

Constraint Qualifier

Goal

Control Constraint Descriptor

Acceptable Error

. . .

Detail Level

Classification

. . .

Detection

Quality Criteria

Performance Criteria

On Boundary

. . .

Accuracy

Occurrences

Classify Image/Video

Detect Presence

Count Occurrence

best_compromise
reliability
robustness

memory
processing_time

.
.
.

prefer_false_alarm_than_miss
prefer_miss_than_false_alarm

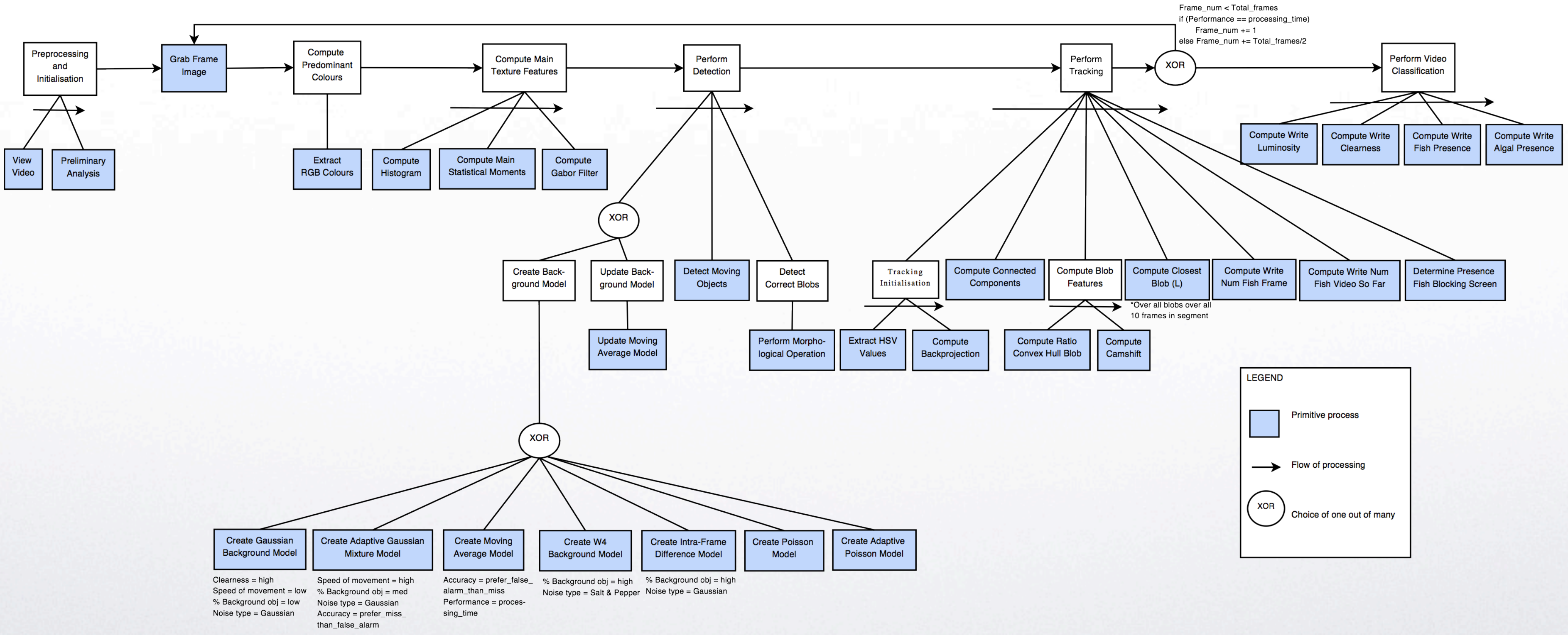
all_occurrences
at_least_one_occurrence
one_occurrence

classify_video

detect_presence_fish
detect_presence_coral

count_fish

High level goals mapped to VIP tools (capabilities) via HTN Planning



Preprocessing and Initialisation

View Video

Preliminary Analysis

Grab Frame Image

Compute Predominant Colours

Extract RGB Colours

Compute Main Texture Features

Compute Histogram

Compute Main Statistical Moments

Compute Gabor Filter

Perform Detection

XOR

Create Back-ground Model

Update Back-ground Model

Update Moving Average Model

Detect Moving Objects

Detect Correct Blobs

Perform Morpho-logical Operation

XOR

Create Gaussian Background Model

Clearness = high
Speed of movement = low
% Background obj = low
Noise type = Gaussian

Create Adaptive Gaussian Mixture Model

Speed of movement = high
% Background obj = med
Noise type = Gaussian
Accuracy = prefer_miss_than_false_alarm

Create Moving Average Model

Accuracy = prefer_false_alarm_than_miss
Performance = processing_time

Create W4 Background Model

% Background obj = high
Noise type = Salt & Pepper

Create Intra-Frame Difference Model

% Background obj = high
Noise type = Gaussian

Create Poisson Model

Create Adaptive Poisson Model

Perform Tracking

Tracking Initialisation

Extract HSV Values

Compute Backprojection

Compute Connected Components

Compute Blob Features

Compute Ratio Convex Hull Blob

Compute Camshift

Compute Closest Blob (L)

*Over all blobs over all 10 frames in segment

Compute Write Num Fish Frame

Compute Write Num Fish Video So Far

Determine Presence Fish Blocking Screen

XOR

```
Frame_num < Total_frames
if (Performance == processing_time)
     Frame_num += 1
else Frame_num += Total_frames/2
```

Perform Video Classification

Compute Write Luminosity

Compute Write Clearness

Compute Write Fish Presence

Compute Write Algal Presence

LEGEND

🟦 Primitive process

→ Flow of processing

XOR Choice of one out of many

2. Optimal Job Scheduling

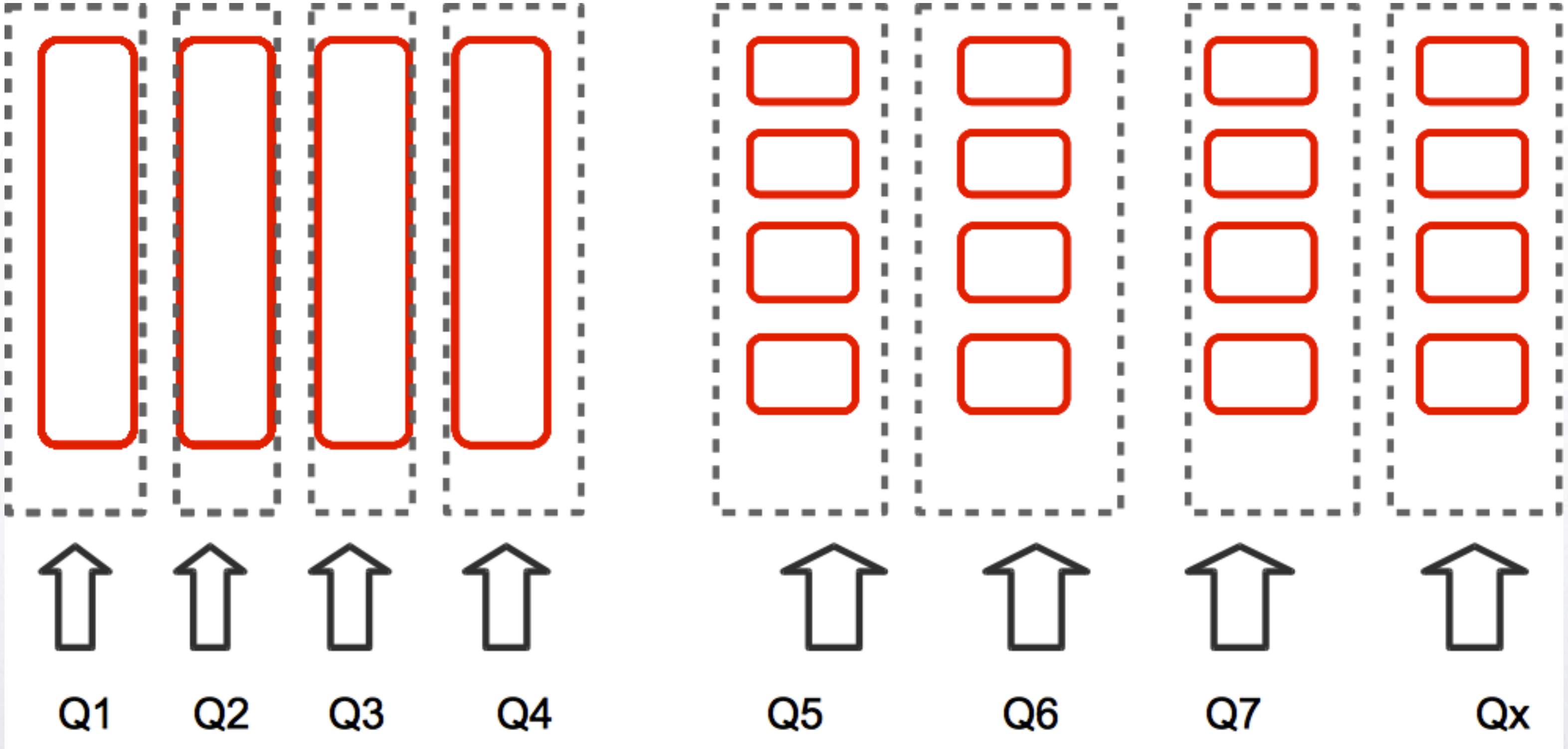
- Facilities: "VMware On-Demand" platform, 48 CPUs, 144G RAM
- Distribution and parallelisation considered
- Is it worth parallelising VIP tasks. If so, which ones?

Module	Ideal	Expected	Worst-case
Fish detection	real-time	20mins	60mins
Fish description	1s per fish	5s per fish	10s per fish
Fish classification	0.1s per fish	1s per fish	2s per fish
Fish clustering	0.1s per fish	1s per fish	2s per fish

Scheduling Options

- Run VIP modules in the pure and native VM Server : shared memory architecture (a pure 48 CPUs SMP system : x86_64)
- Run VIP modules in one VM system : shared memory architecture (a virtual SMP system : x86_64)
- Run VIP modules in several VM systems : individual and distributed memory architecture (virtual Cluster system), similar to the real PC Cluster. (MPI / PVM).

Parallelise Whole Tasks



3. Live Stream Processing

- Live streams pose changes in the data that come in within a single workflow execution
- How to model an adaptive workflow solution that can manage such scenario
- Consider Cloud solutions, e.g. Hadoop for dynamic scalability?

Data Streaming *vs.* Job-scheduling

- Is it worth looking into data streaming?
- Minimises intermediate data which is useful for VIP domain
- Job-scheduling also has ways of reducing data movement, e.g. clustering
- Job scheduling and then data streaming?

TODOs

- Ontologies w.r.t. queries (goals), fish and video descriptions, processes and capabilities to be defined
- Investigate what data flows could benefit from shared memory *vs.* file sharing
- Investigate processor/memory and control capabilities
- Develop simulated F4K system to evaluate different allocation strategies
- Decompose 20 Qs into tasks
- Investigate multiprocess scheduling, e.g. Beowulf

Demo: SWAV

Related Projects

- Pan-STARRS http://pan-starrs.ifa.hawaii.edu/public/
- Admire http://www.admire-project.eu