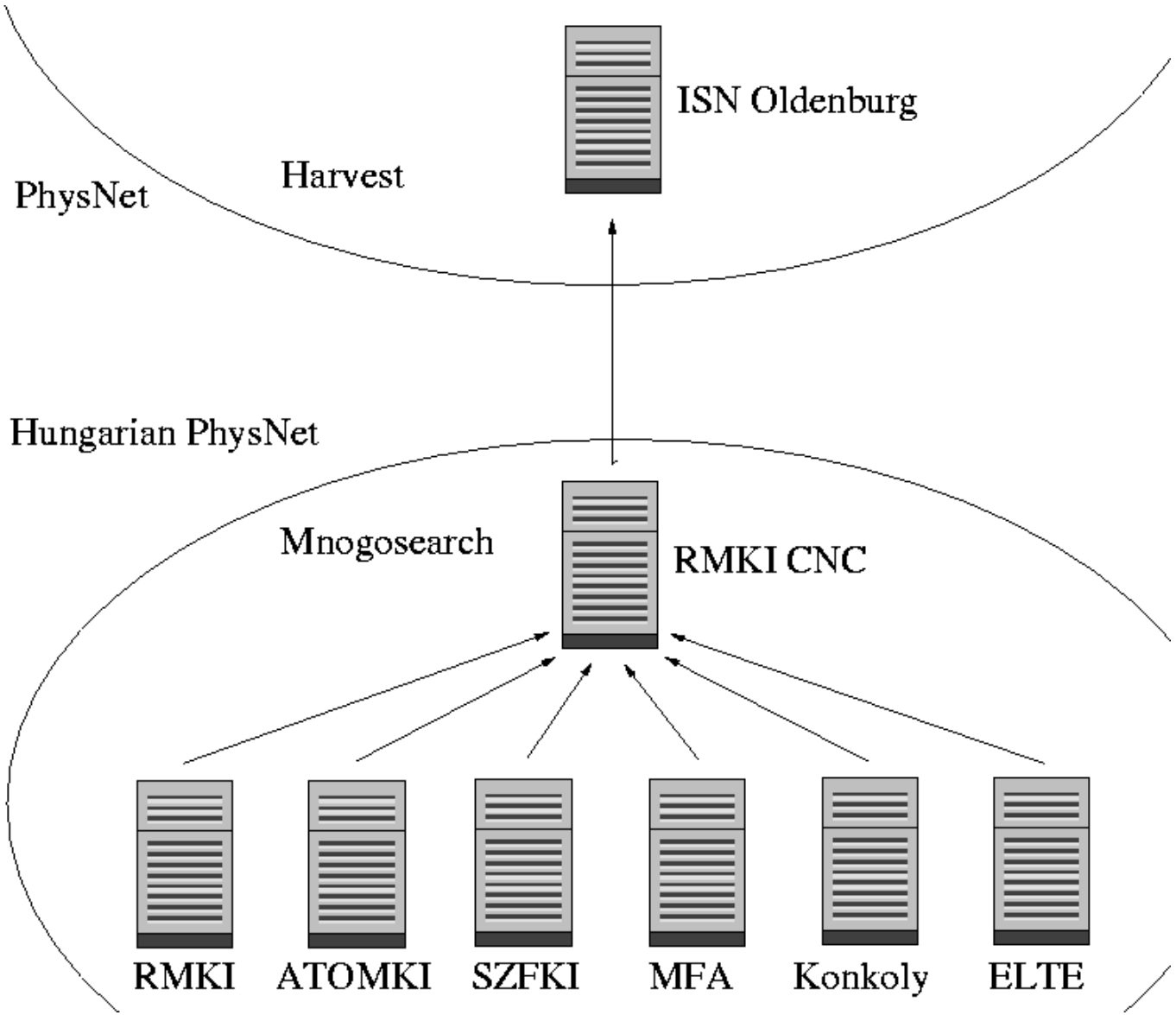
Search Engine Tree in Hungarian PhysNet

KFKI RMKI CNC
Budapest, Hungary
Kati Szalay <szalay@sunserv.kfki.hu>
József Kadlecsik <kadlec@sunserv.kfki.hu>

PhysNet cooperation in Hungary

- KFKI Research Institute for Particle and Nuclear Physics (KFKI RMKI)
- ATOMKI Research Institute of Nuclear Research
- Research Institute for Solid State Physics and Optics (SZFKI)
- Research Institute for Technical Physics and Material Sciences (MFA)
- Konkoly Observatory
- Eötvös Loránd University of Sciences (ELTE)

Planned search engine tree



MnoGoSearch

- SQL database backend
  - MySQL
- searchd
  - several search databases can be merged

mySQL database replication

- Databases are not overlapping

- mySQL internal replication support: masters -> slaves
- via dumpfile
  - master: mysql -e "SELECT ..." > outfile
  - rsync
  - slave: load_data
- slave: SELECT <updated>, SELECT <inserted>

Words storage modes

- single: one table
- multiple: 13 tables depending on word length
- crc: 32 bit integer word IDs stored
- crc multi: multiple tables with crc32 IDs
- cache: word index is stored on disk

URL table

rec_id

docsize

next_index_time

last_mod_time

referer

crc32

url

...

urlinfo table

url_id	i.e rec_id from url table
sname	(content-type, title, body, meta tags, ...)
sval	

dict table

url_id i.e rec_id from url table

word

intag

SOIF

- Summary Object Interchange Format:

SOIF = OBJECT SOIF | SOIF

OBJECT = @ TEMPLATE-TYPE { URL
ATTR-LIST }

ATTR-LIST = ATTRIBUTE ATTR-LIST | ATTRIBUTE

ATTRIBUTE = ID{VALUE-SIZE} DELIM VALUE

DELIMITER = ":<TAB>"

Common SOIF Attributes

- File-Size: url.docsize
- Gatherer-Host|Name|Port|Version
- Refresh-Rate, Time-to-Live
- Update-Time: url.next_index_time - Refresh-Rate
- Last-Modification-Time: url.last_mod_time
- MD5: url.crc32
- Type: urlinfo.sname = content-type
  where urlinfo.url_id = url.rec_id

Common SOIF Attributes, cont.

- Title: urlinfo.sname = title
- Keywords: urlinfo.sname = meta.keywors
- Partial-Text: dict.word
- URL-References: no mnoGoSearch equivalent!

The grand plan

- Compare effectiveness, stability, robustness of different mySQL database replication methods
- Develop mnoGoSearch - Harvest gateway
- Deploy hardware, software
- Start the service