Hyperscan in Rspamd

When PCRE is not enough

Problem definition

- Need to match many regular expressions for the same data:

```
len: 610591, time: 2492.457ms real, 882.251ms virtual
regexp statistics: 4095 pcre regexps scanned, 18 regexps matched, 694M bytes scanned using pcre
```

- No need to capture patterns
- Need to match both UTF8 and raw data
- Need to support Perl compatible syntax

Naive solution

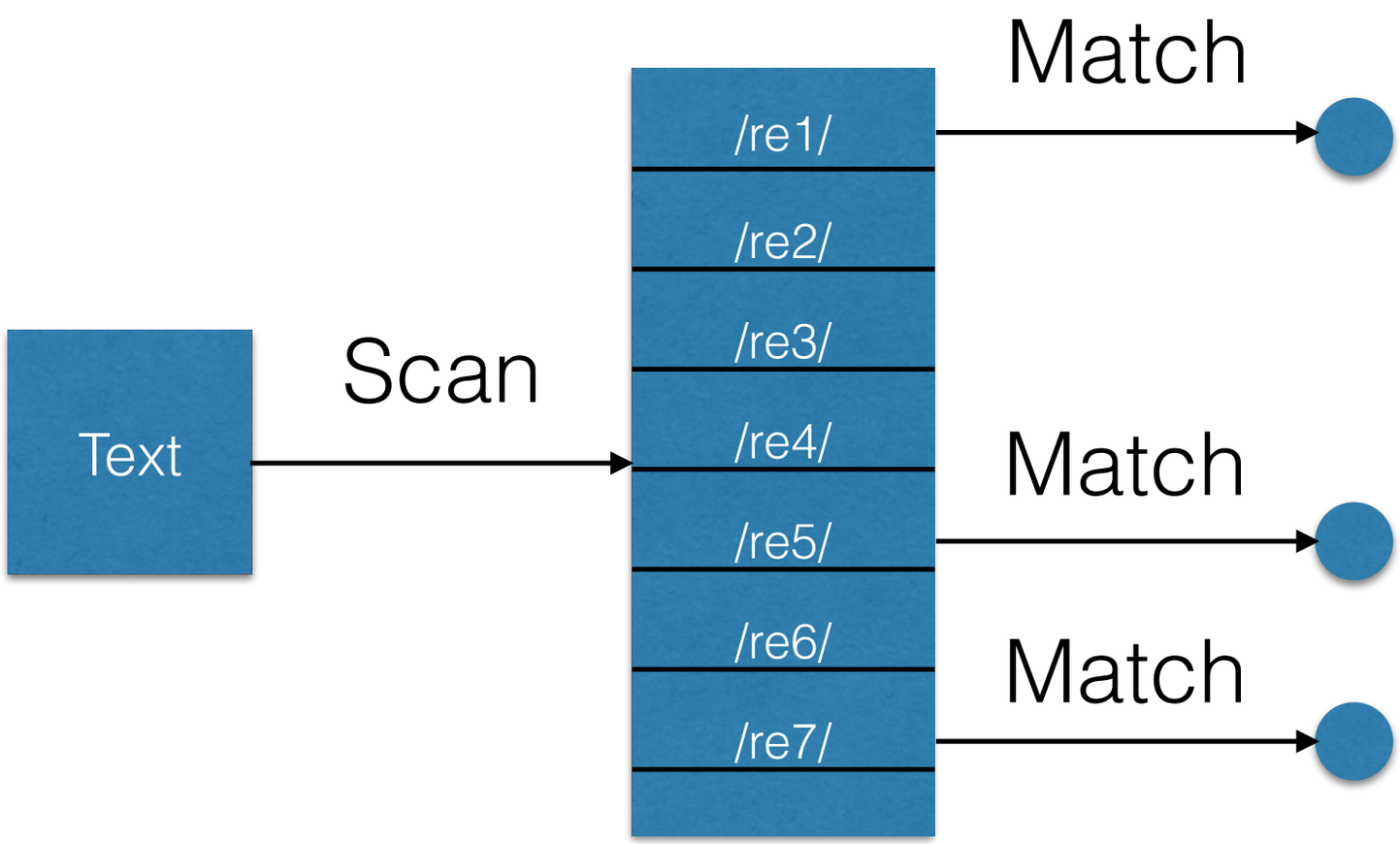
- Join all expressions with ‘|’:

```
/abc/ + /cde/ => /(?:abc)|(?:cde)/
```

- Won’t work:

```
/abc/ + /a/ => /(?:abc)|(?:a)/ -> the second rule won’t match for ‘abc’
```

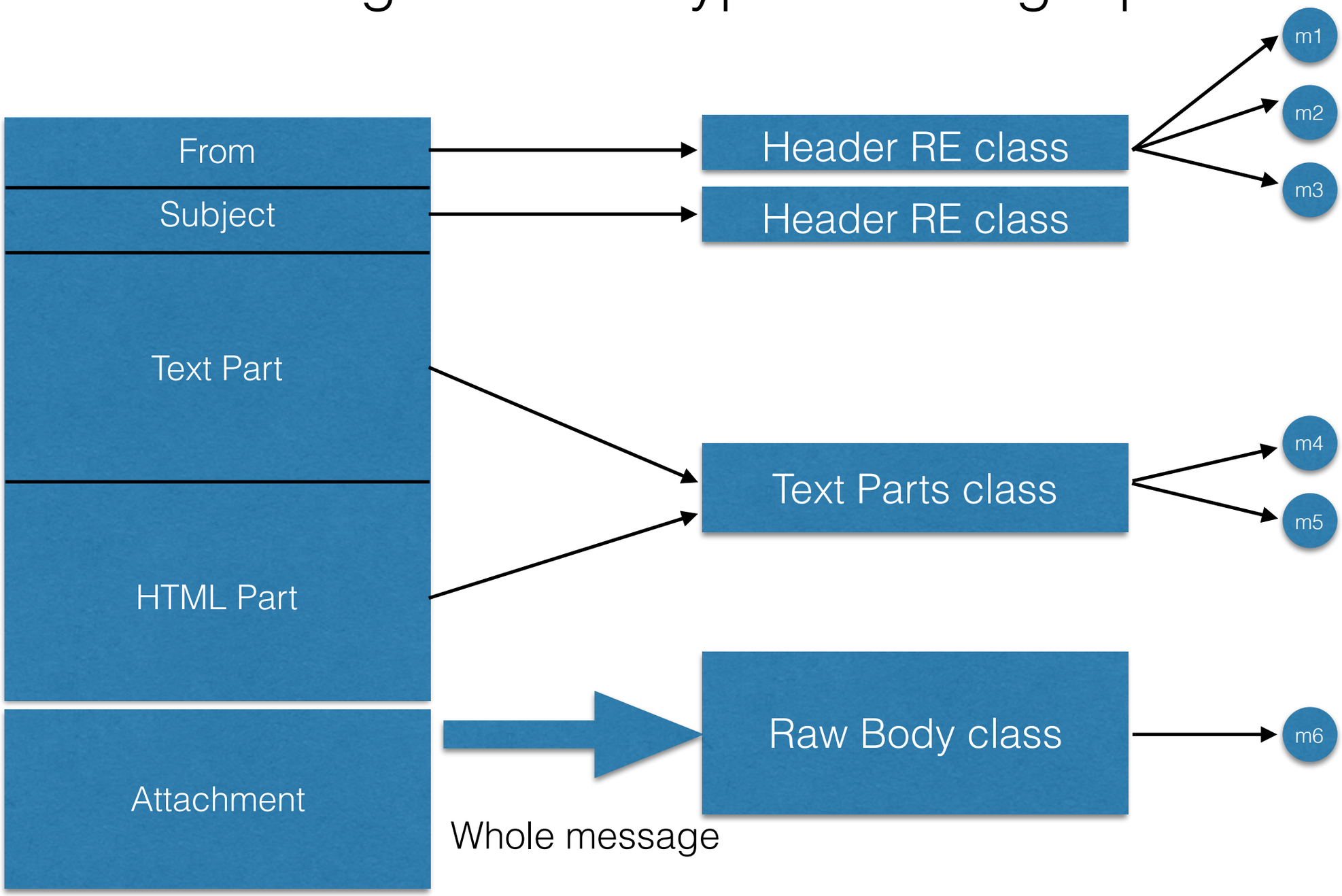
Hyperscan



Rules and Hyperscan

MIME message

Hyperscan regexps



Problems with Hyperscan

- Slow compile time (like 100 times slower than PCRE+JIT)
- Many PCRE features are unsupported (lookahead, look-behind)
- Is not packaged for the major OSes (requires relatively fresh boost)

Lazy hyperscan compile

Main

Compile to disk

HS compiler

Compiled class

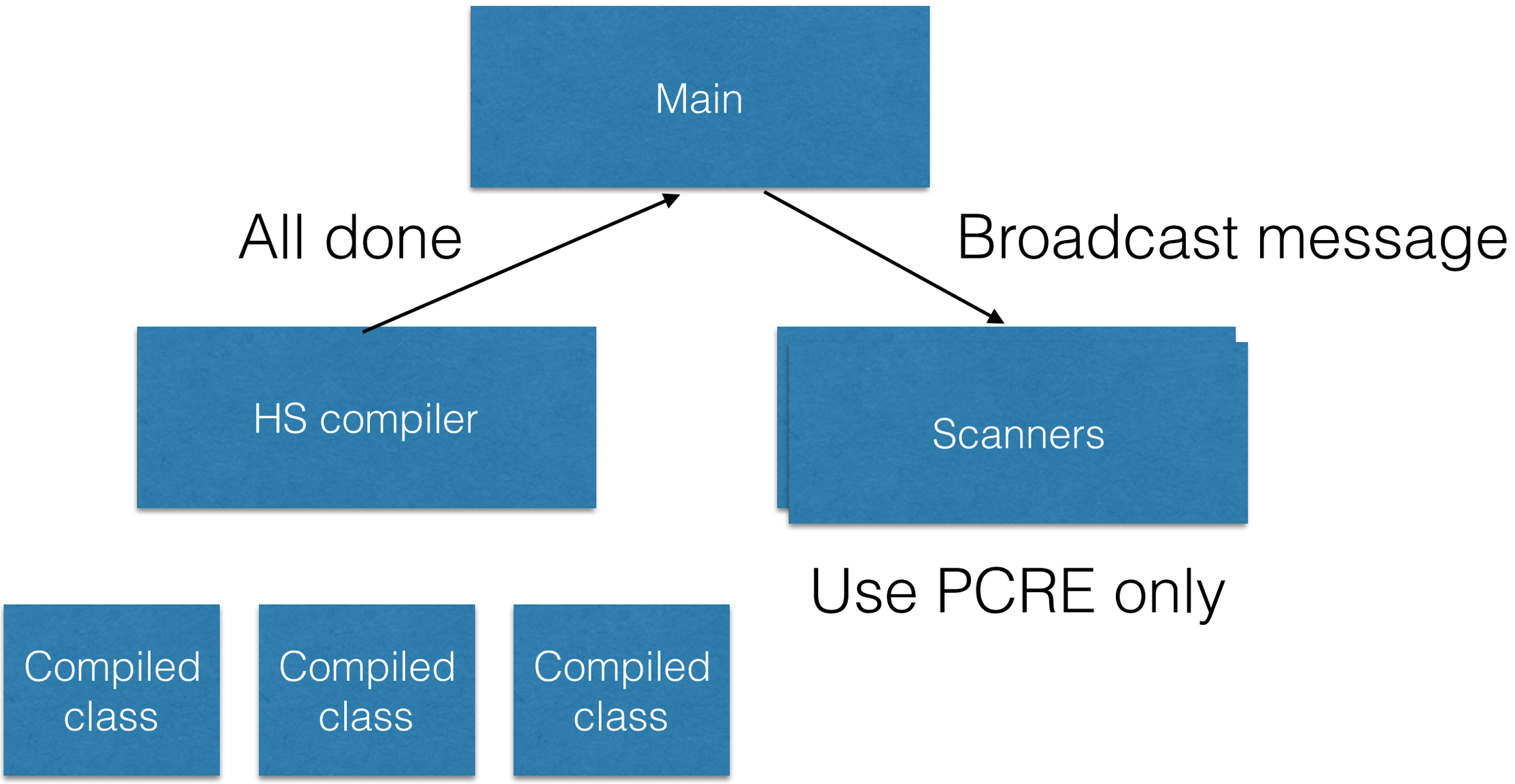
Compiled class

Compiled class

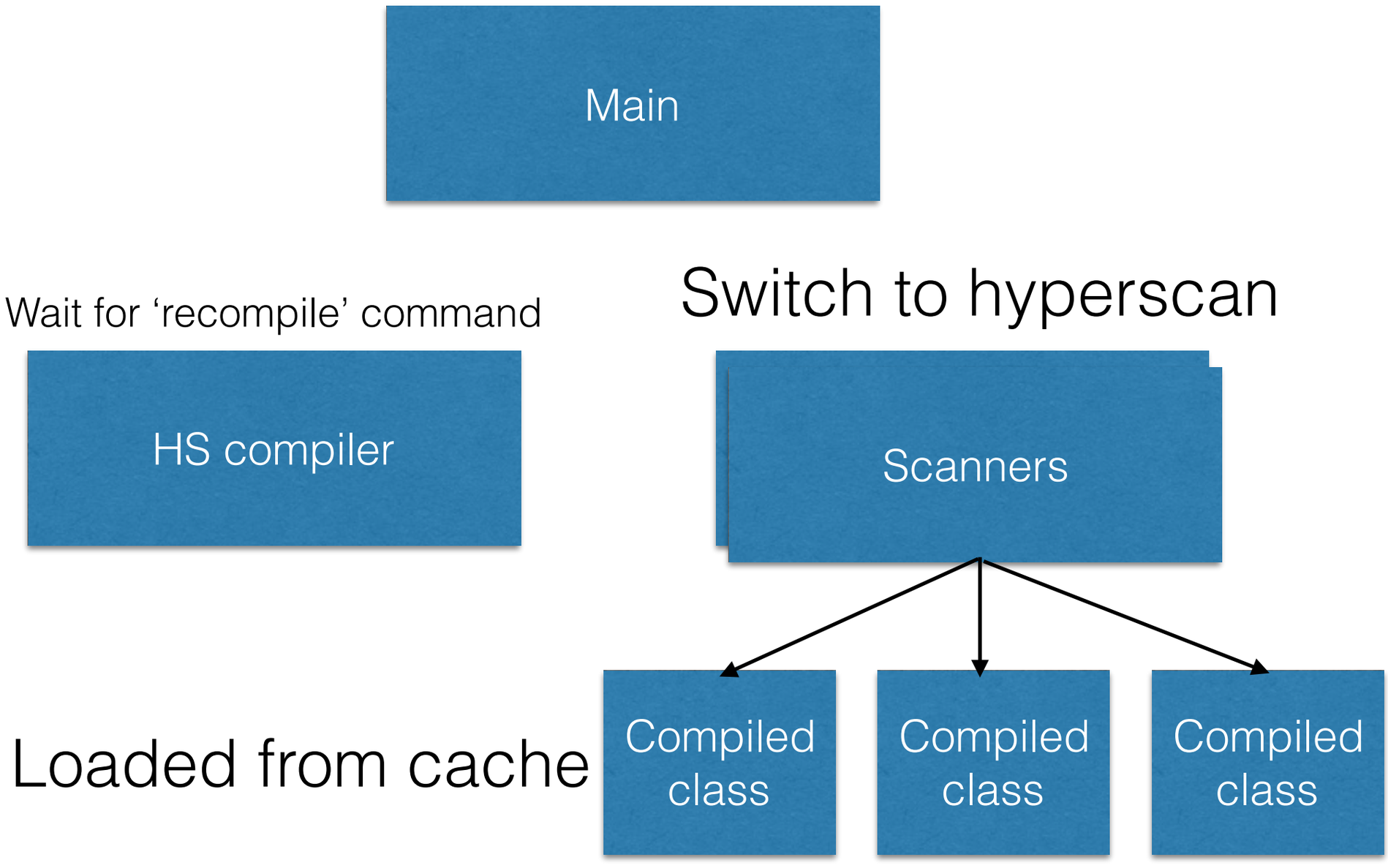
Scanners

Use PCRE only

Lazy hyperscan compile



Lazy hyperscan compile



Lazy hyperscan compile

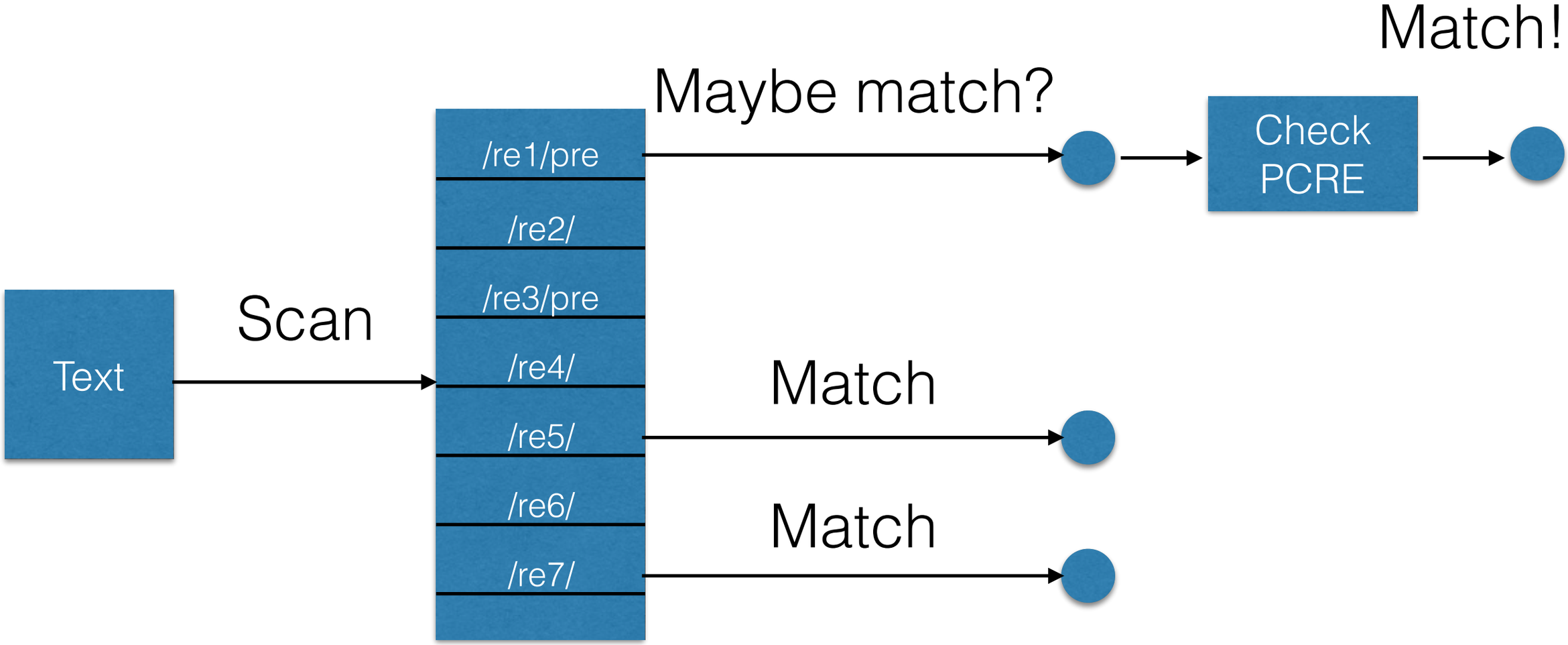
- Does not affect the starting time
- If nothing changes then scanners loads pre-compiled expressions on startup (using crypto hashes)
- Recompile command is used to force recompilation process:

```
rspamadm recompile
```

Unsupported features

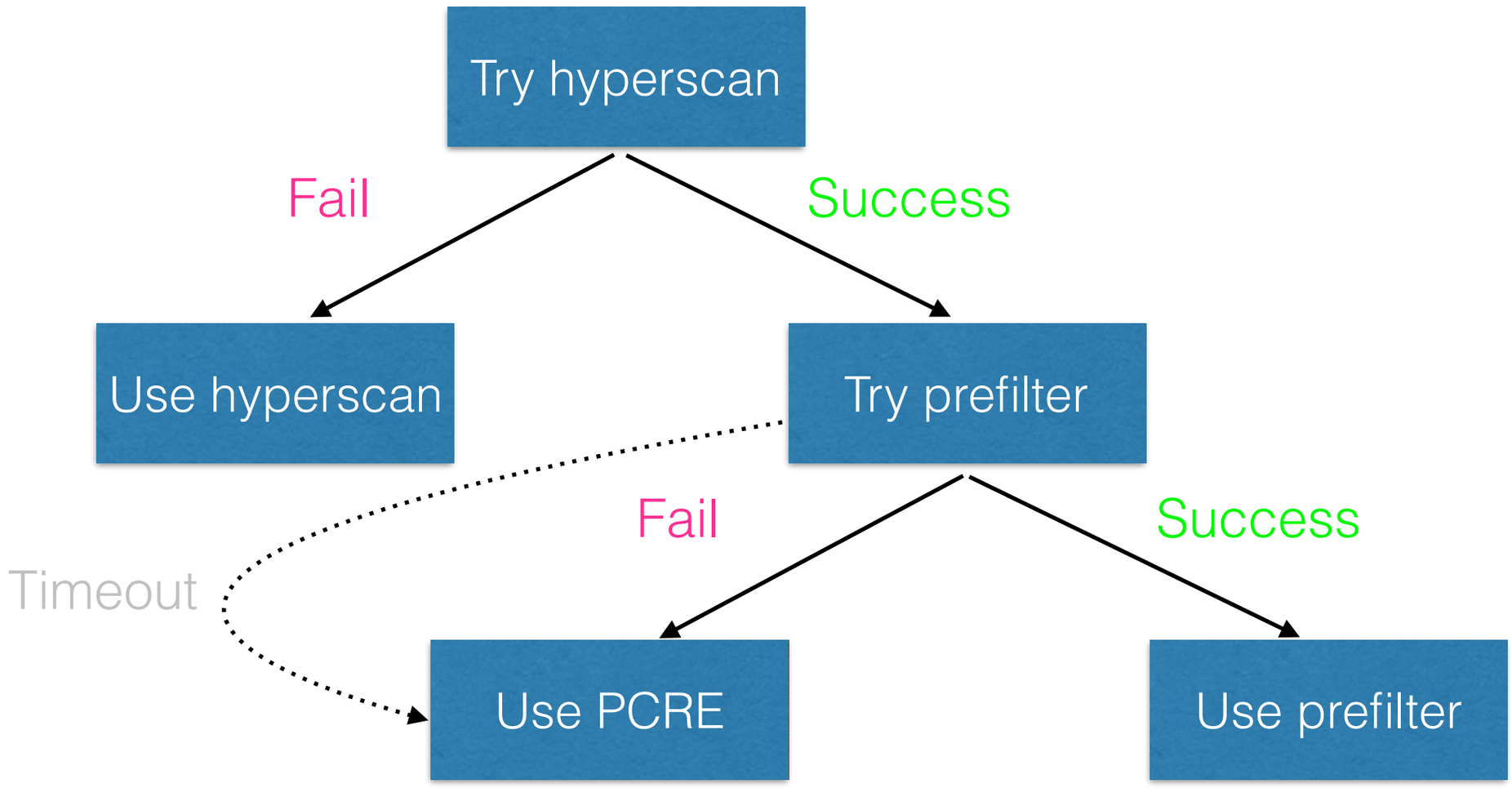
Pre-filter mode

Replaces unsupported constructs with "weaker" supported ones, guaranteeing a match superset:



Unsupported features

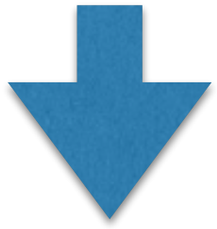
Approximation



Results

- Use Hyperscan for the vast majority of RE rules:

len: 610591, time: 2492.457ms real, **882.251ms** virtual
regexp statistics: **4095** pcre regexps scanned, 18 regexps matched, **694M** bytes scanned using pcre



len: 610591, time: 654.596ms real, **309.785ms** virtual
regexp statistics: **34** pcre regexps scanned, 41 regexps matched, **8.41M** bytes scanned using pcre, **9.56M** bytes scanned total

Results

- The default rules all compile with Hyperscan without prefiltering mode:

```
compiled 235 regular expressions to the hyperscan tree
loading hyperscan expressions after receiving compilation notice
hyperscan database of 235 regexps has been loaded
```

- Works for most of the SA rules (some are prefiltering):

```
hyperscan database of 4118 regexps has been loaded
```

- Speeds up both small and large messages

Questions?

Vsevolod Stakhov
https://rspamd.com
https://01.org/hyperscan