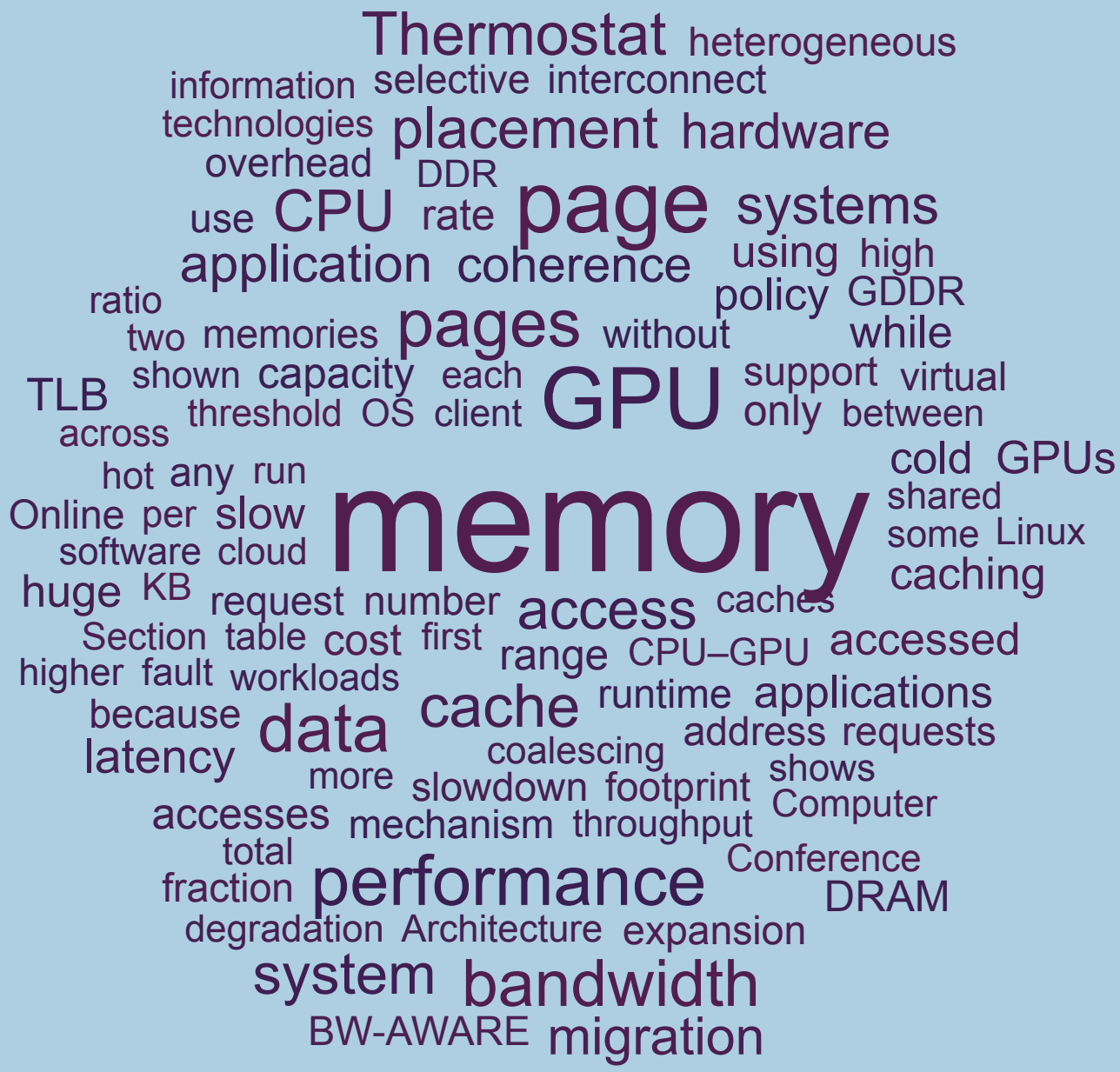
Thermostat
heterogeneous
information
selective
interconnect
technologies
placement
hardware
overhead
DDR
use
CPU
rate
page
systems
application
coherence
using
high
ratio
policy
GDDR
two
memories
pages
without
while
shown
capacity
each
GPU
support
virtual
TLB
threshold
OS
client
only
between
across
cold
GPUs
hot
any
run
memory
shared
Online
per
slow
some
Linux
software
cloud
caching
huge
KB
request
number
access
caches
Section
table
cost
first
range
CPU–GPU
accessed
higher
fault
workloads
because
data
cache
runtime
applications
latency
coalescing
address
requests
more
slowdown
footprint
shows
accesses
mechanism
throughput
Computer
total
Conference
fraction
performance
DRAM
degradation
Architecture
expansion
system
bandwidth
BW-AWARE
migration